Information Technology: Research and Development (1982) *1* (177–187)

ONLINE IDENTIFICATION OF WORD VARIANTS AND ARBITRARY TRUNCATION SEARCHING USING A STRING SIMILARITY MEASURE

G. E. FREUND AND P. WILLETT*

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, UK

(Received 19 October 1981, revised 2 February 1982)

ABSTRACT

A method is described for the interactive identification of word variants, e.g., grammatical variants or misspellings, based on the numbers of *n*-grams common to pairs of words. Experiments with query terms from two document test collections suggest that the method can identify a large proportion of the word variants present in a file of documents, without retrieving an unacceptable number of non-related terms. The use of the technique for arbitrary truncation searching, and its advantages over conventional online conflation methods are discussed.

1. METHODS FOR THE CONFLATION OF WORD VARIANTS

The popularity of online bibliographic retrieval systems is evidenced by the fact that there were, by the end of 1980, almost 200 bibliographic databases with more than 65 million references available for interactive searching (Hall and Brown, 1981). However, current online retrieval systems exhibit several limitations which restrict their utility for retrieving the documents relevant to some query: examples of these defects include the limited browsing facilities, the confusing range of retrieval languages and system features, and the frequent lack of a ranking procedure for search output. Among the greatest problems are the variations, inconsistencies, and errors which are encountered by the searcher in the terms used for database indexing. There are many types of word variant, the most common being:

1. Spelling errors.
2. Valid alternative spellings, e.g., NEIGHBOUR and NEIGHBOR in English and American.
3. Alternative forms of multi-word concepts, e.g., ONLINE, ON-LINE, and ON LINE.
4. Transliteration problems, e.g., TCHEBYSHEFF and CHEBYSHEFF.

* To whom correspondence should be addressed.

5. The effect of affixes, e.g., SINGLE and SINGULAR, or UNIFY and REUNIFICATION.
6. Alternative forms of abbreviation as in APPROX and APPROXN.

Many methods have been described in the literature for the conflation of certain of these types of word variant (Hall and Dowling, 1980). Thus Paice (1977) has given a set of rules for the conversion of English and American spellings of the same word to a common form while Aldefeld *et al.* (1980) have described the conflation of phonetically similar surnames, e.g., RODGERS and ROGERS, as a component of an automatic telephone directory system. A comparative evaluation of the characteristics of a range of word abbreviation schemes was reported by Bourne and Ford (1961), and Pollock (1977) has described the routines used at Chemical Abstracts Service for the systematic generation of abbreviations.

An important group of procedures is formed by the stemming algorithms which have been used as a means of conflating grammatical variants in natural language understanding (Cercone, 1978), literary analysis (Raben and Lieberman, 1976), and information retrieval systems (Lovins, 1968; Lennon *et al.*, 1981). A stemming algorithm is a computational procedure which strips the affixes from words so that only the word root need be stored for subsequent processing; most algorithms remove only suffixes so that word variants are conflated to their common stem.

Ignorance, faulty keyboarding, noisy transmission lines, and OCR devices are examples of the means by which spelling errors may be introduced into machine-readable texts. The most common types of error involve the transposition of two characters, the insertion or deletion of a single character, or the erroneous substitution of one character for another (Damerau, 1964; Ullman, 1977), and a large body of research has been directed to the identification and subsequent correction of such misspellings (Blair, 1960; Davidson, 1962; Morgan, 1970; James and Partridge, 1976; Peterson, 1980). However, despite intensive efforts by the producers of bibliographic databases to eliminate errors (Zamora, 1980), Bourne (1977) found that, on average, no less than 10.8 per cent of the index terms in 11 online databases were misspelled in some way.

The primary means of conflation provided by the online vendors are right-hand truncation and the ability to inspect specific parts of the inverted file where related terms might appear, as in the NEIGHBOR and EXPAND commands of the ORBIT and DIALOG search systems. These commands enable the user to display a certain number of the dictionary terms which are immediately adjacent to a given query term. In this way, the searcher can identify variants of the query term, provided that the alternative forms appear near the query term in the alphabetical sequence; thus terms like NONLINEAR or the misspelling LYNEAR might well be omitted in a search involving the term LINEAR.

2. WORD CONFLATION USING THE CONCEPT OF STRING SIMILARITY

The procedures described above are generally unsuited to the conflation of all possible types of word variant, and they also exhibit specific defects even in the chosen area of application. Thus Lovins (1971) has analysed errors in stemming algorithms, and Toussaint (1978) has considered the effects of context on the accuracy of spelling correction procedures.

A very different approach to the problem of word variants was devised by

Adamson and Boreham (1974) who suggested that 'the character structure of a word is so related to its semantic content as to make this a useful basis for automatic classification of words'. Their analysis consisted of representing each word in a dictionary by a list of its constituent *n*-grams, i.e., strings of *n* adjacent characters, and then using these lists as the basis for calculating a quantitative measure of similarity between pairs of words. Thus the words CONSTRUCT and DESTRUCTION may be represented by the lists of digrams

CO ON NS ST TR RU UC CT

and

DE ES ST TR RU UC CT TI IO ON

from which it can be seen that the words have 6 digrams in common. It is, of course, possible to determine similarity using the occurrence of single characters as the attributes which are to be compared. Words with the same root would indeed be identified in this way but many other words would be identified as matching when they did not, in fact, share a common root. Such erroneous conflations will decrease if larger *n*-grams are considered; however long substrings, such as tetragrams, may mean that short common roots are missed, e.g., MIX and PREMIXED, or that spelling errors may lead to a large reduction in the number of common *n*-grams even with related terms. Once the number of common *n*-grams has been identified, a quantitative measure of similarity may be obtained by evaluating a similarity coefficient. Writing $NG(W1)$ and $NG(W2)$ for the numbers of *n*-grams in two words $W1$ and $W2$, and $NG(C)$ for the number of *n*-grams in common, examples of common similarity coefficients are the Dice coefficient $2NG(C)/(NG(W1)+NG(W2))$, and the Overlap coefficient $NG(C)/\min(NG(W1),NG(W2))$. Discussions of textual similarity measures are given by Alberga (1967), Stalker (1978), and Findler and Van Leeuwen (1979).

Adamson and Boreham (1974) used inter-word similarity coefficients as a basis for a clustering of words and found that intuitively reasonable groups of words were formed. They then clustered a small group of mathematical titles on the basis of their constituent digrams and suggested that this could form the basis for a document retrieval system; Willett (1979), however, found that the procedure gave poor results in experiments with the Cranfield test collection. Lennon *et al.* (1981) used a related approach for the automatic expansion of terms in a query. A dictionary was created containing all of the terms occurring in the 1396 titles of the Cranfield document test collection and these terms were represented by the lists of constituent digrams or trigrams. Each of the terms in the 225 queries was then similarly represented and matched against each of the words in the index term dictionary (hereafter abbreviated to ITD). Index terms with a similarity coefficient greater than some threshold value were considered to be variants of the query term and automatically added to the query: the expanded queries were then used for searching the file of documents in the normal way.

This approach can handle all of the types of variant discussed in the first section of this paper. For example, when given the query term CONDUCTOR, the query might be expanded to include the following, *inter alia*,

1. Related words with different affixes, e.g., CONDUCTORS, CONDUCTION,

SUPERCONDUCTORS, PHOTOCONDUCTANCE, and SEMICONDUCTOR.
2. Different word forms, e.g., SUPERCONDUCTOR and SUPER-CONDUCTOR.
3. Misspellings, e.g., CONDUTOR and COMDUCTOR.

In a series of retrieval experiments, this similarity procedure was found to result in a level of retrieval effectiveness which was at least comparable with the levels obtained from the use of a range of conventional stemming algorithms (Lennon *et al.*, 1981). The technique does, however, have several limitations:

1. Words may be added to the query which are not, in fact, variants of the query term, although the query and index terms have a high similarity coefficient, e.g., RUNNING and CUNNING, or CONSUMPTION and CONSTRUCTION.
2. Words may be added to the query which are indeed related to the sought word but which are not relevant to the search owing to the particular affixes present, e.g., CONSTRUCTION and DESTRUCTION, or ABILITY and DISABILITY.
3. Query terms may have more than one meaning. Thus most of the terms listed above as being related to CONDUCTOR would not be applicable to a musical query.

For these reasons, a system has been developed in which possibly related terms are presented at a terminal in real time, and it is up to the searcher to decide whether a particular term should, or should not, be added to the query.

3. ONLINE QUERY EXPANSION

The procedure used by Lennon *et al.* for automatic query expansion involved the matching of a query term against each and every member of the ITD. This is much too slow for an interactive system and we have accordingly adopted the efficient inverted file structure described by Noreault *et al.* (1977). Inverted file systems are widely used for Boolean document retrieval but Noreault *et al.* showed how they could also be used for the calculation of similarity coefficients and for the production of ranked output. Their algorithm involves the vectorial addition of the lists of document numbers from the inverted file which correspond to the terms in a query. The addition results in the calculation of the number of terms in common between each document and the query; knowing the number of terms in the document and in the query, a similarity coefficient may readily be calculated. A detailed discussion of implementation techniques for the vectorial addition has been given by Willett (1981) in the context of automatic document classification.

The system described here involves the creation of an inverted file to the *n*-grams in the ITD: digrams and trigrams were used in our experiments. Each entry in the inverted file consists of:

1. An *n*-gram.
2. A pointer to a list which contains the term numbers for each occurrence of that *n*-gram in a term in the ITD.

When a query term is submitted to the system, it is broken down into its constituent *n*-grams, the appropriate lists identified in the inverted file, and the lists added so as to identify the number of *n*-grams common to the query term and to each of the words in the ITD. This information is used for the calculation of the Dice similarity coefficients, and the high-ranking terms from the ITD are then displayed on a terminal for consideration as potential additional search terms.

The effectiveness of the technique was evaluated using a suite of PASCAL programs which were run on the University of Sheffield PR1ME 750 minicomputer. These programs create the inverted files, match the query against the file, and then display possible search terms: a full description of the system and of the experiments described below is given by Freund (1981).

Two sets of documents and query terms were used:

1. 121 query terms and 5220 document terms from the Evans document test collection.
2. 151 query terms and 11997 document terms from the Vaswani document test collection.

A displayed index term was considered to be related to a query term if it

1. Had the same basic character structure as the query term and was semantically related to it, e.g., *SECONDARY* and SECOND, *CRYSTAL* and CRYSTALLINE, RECRYSTALLIZATION, SEMICRYSTALLINE, or
2. Was an unmistakable misspelling of the query term which could not have arisen from the misspelling of another word, e.g., *CONDUCTOR* and CONDDUCTOR, *ATTENUATION* and ATTENTUATE.

(In each case, the query term has been italicized.) Conversely, a displayed term was considered to be unrelated if it

1. Had the same basic character structure as the query term but was not semantically related to it, e.g., *STABLE* and ABLE, TABLE, *DEFECT* and DEFLECT.
2. Was semantically related but had been retrieved through a similarity in affixes, rather than via a similarity in the word root, e.g., *LIQUID* and FLUID, *PRIMARY* and SECONDARY.
3. Could have been a misspelling of more than one word, e.g., *PULSE* and PULE (which could also have been derived from PULL).

It was found that very large numbers of non-related words were retrieved at low similarity levels and hence when a query term was submitted to the system, a threshold similarity was specified so that only those index terms with a similarity greater than the threshold were displayed. Threshold values from 0.80 down to 0.40 in 0.05 intervals were tested.

An inspection of the inverted files showed a well-marked Zipfian distribution of *n*-gram usage. Thus, of the 628 digrams in the Vaswani inverted file, 342 of them occurred in less than 50 words; conversely the digram 'S ', i.e., a terminal S, occurred in no less than 2519 words. Obviously, words containing common affixes such as -ATION, -ING, or RE- may lead to the retrieval of a large number of non-related words, especially if the affixes are connected to a short root, or if more than

one of them occur in a word. For this reason, common affixes were removed from query terms before they were input to the retrieval system. As an example, the digrams from the query term ATTENUATION retrieved 8 related and 308 non-related words at a similarity threshold of 0.40, whereas the truncated query term ATTENUAT retrieved the same 8 related words but only 37 non-related terms. It is obviously not possible to delete common affixes in some cases, e.g., ION in IONIZATION or ING in SINGING, but is felt to be reasonable to ask the user of such a system to avoid common affixes wherever possible. Two alternative possibilities are:

1. To use a similarity function in which the variant frequencies of occurrence of individual *n*-grams are taken into account. In this case, a match on an infrequently occurring *n*-gram would count for more than a match on a common one, and this would tend to reduce the effect of the common *n*-grams. (A similar approach has been used to counter the effect of highly posted query terms in document retrieval systems: Sparck Jones, 1972.)
2. To use variable-length *n*-grams, chosen so as to occur approximately equifrequently (Lynch, 1977), rather than the fixed-length, variable-frequency *n*-grams used here.

```
Query term  :  LINEAR

Index terms :  CURVILINEAR, LINE, LINEAR,

    LINEARISED, LINEARIZATION, LINEARIZED,

    LINEARLY, LINES, LINED, NEAR, NONLINEAR,

    RECTILINEAR

Query term  :  STABILIZE

Index terms :  DESTABILIZING, INSTABILITY, STABILISE,

    STABLISED, STABILISER, STABILISING,

    STABILITY, STABILIZATION, STABILIZED,

    STABILIZER, STABILIZING
```

FIG. 1. Examples of retrieved index terms using trigrams with a threshold similarity of 0.50

For each query term, truncated if necessary, the number of related, R_t, and non-related, N_t, terms retrieved at some threshold t were noted. The average results obtained are shown in Tables 1 and 2, and some typical searches in Figure 1. (It should be noted that the searches presented in Figures 1 and 2 correspond to actual searches using the Evans or Vaswani ITDs and different sets of words would be retrieved using other dictionaries.) Several points should be made about these results:

1. The use of digrams leads to much larger numbers of non-related terms, especially at the lower similarity thresholds; conversely, the numbers of index terms retrieved using trigrams is quite acceptable for rapid visual inspection at a terminal.

Table 1. Mean numbers of related, R_t, and non-related, N_t, terms retrieved over a range of similarity thresholds, t, for the Evans ITD

	Digrams		*Trigrams*	
t	R_t	N_t	R_t	N_t
0.80	1.2	0.1	1.0	0.0
0.75	1.7	0.1	1.4	0.0
0.70	2.6	0.2	2.1	0.0
0.65	3.0	0.5	2.5	0.1
0.60	3.6	1.2	2.9	0.4
0.55	4.3	2.6	3.5	0.7
0.50	5.0	5.9	4.2	1.4
0.45	5.4	9.8	4.7	2.0
0.40	5.8	23.9	5.3	4.4

2. The figures quoted are mean values and some of the digram searches resulted in the retrieval of very large numbers of words.

Table 2. Mean numbers of related, R_t, and non-related, N_t, terms retrieved over a range of similarity thresholds, t, for the Vaswani ITD

	Digrams		*Trigrams*	
t	R_t	N_t	R_t	N_t
0.80	1.7	0.0	1.3	0.0
0.75	2.9	0.1	2.0	0.0
0.70	4.3	0.4	3.2	0.1
0.65	5.2	0.9	4.1	0.2
0.60	6.3	2.3	5.1	0.6
0.55	7.5	5.0	6.1	1.2
0.50	8.4	11.3	7.2	2.3
0.45	8.9	18.9	7.7	3.4
0.40	9.9	46.9	8.7	7.9

3. The figures quoted reflect only the precision of the system, i.e., the proportion of the retrieved terms which were related according to the definition given above. The recall of the searches could be determined exactly only by a manual examination of the entire ITD for each query term. However, the recall at some threshold t can be estimated with little error by $R_t/R_{0.40}$, where $R_{0.40}$ is the number of related terms retrieved using digrams and with a tnreshold of 0.40, since examination of the ITD for a few randomly selected query terms showed that very few related terms were in fact present in the file but not retrieved by this parameter setting. Thus, an inspection of the tables shows that the use of trigrams, which retrieved a much higher proportion of related material, does not appear to seriously affect the recall performance of the system.

4. Although the Vaswani ITD is 2.3 times the size of the Evans ITD, the ratio of the numbers of words retrieved at a given threshold from the two ITDs is rather less; however, no systematic experiments were undertaken to test the effect of dictionary size on the number of words retrieved.

The main overhead imposed by the proposed method is the storage required for the n-gram inverted file. Assuming that $n-1$ padding space characters are inserted before and after an m character word, each term will give rise to $m+n-1$ n-grams: thus, taking 8 characters as the average length for an English word, most index terms will give rise to about 10 entries in a trigram inverted file. Hence an ITD of 100000 terms would require about one million words of computer storage to accommodate the trigram inverted file: this amount of storage is well within the capacity of the large floppy and Winchester disk units which are now available for use with microcomputer systems. Such a system would be quite sufficient to store the inverted file, calculate the query–index term similarities, and pass the expanded query over to a mini or mainframe computer for the actual online search.

4. ARBITRARY TRUNCATION SEARCHING IN ONLINE SYSTEMS

Most online retrieval systems offer a truncation facility. This is, however, often restricted to right-hand or embedded truncation, i.e., searching for words with an embedded 'don't care' character, owing to the alphabetical ordering of the inverted file. Searching for word endings, rather than word beginnings as in right-hand truncation searching, is rather more difficult and several possible solutions have been suggested:

1. Search the inverted file sequentially and try to match the query character string to each part of each word.
2. Include in the inverted file all of the reversed words and then, for left-hand truncation, reverse the query term before searching. This will not solve the general problem when search terms are truncated at both ends.
3. Store in the inverted file all of the forms of a word which can be obtained by cyclically shifting one character at a time so that, e.g., the word RING would give rise to the additional entries

   ING R, NG RI, and G RIN.

   If right-hand truncation is available, these rotated forms allow access to any substring in the word.

The use of the system described in Section 3 for left-hand and/or right-hand truncated query terms will retrieve many words containing the desired substring. However, two slight modifications will ensure the rejection of the great majority of the words from the ITD which do not contain the sought word fragment. The modifications involve the use of

1. A similarity coefficient $NG(C)/NG(Q)$ where $NG(C)$ is the number of n-grams common to an index and a query term, and where $NG(Q)$ is the number of n-grams in the query term,

2. A threshold similarity of 1.0 which ensures the retrieval only of those index terms which contain all of the query term *n*-grams.

Table 3. Mean numbers of correctly, $R_{1.0}$, and incorrectly, $N_{1.0}$, identified terms retrieved using truncated query terms

	Digrams		*Trigrams*	
Dictionary	$R_{1.0}$	$N_{1.0}$	$R_{1.0}$	$N_{1.0}$
Evans	7.0	0.3	7.0	0.0
Vaswani	11.4	0.5	11.4	0.0

A set of 41 truncated query terms was put to the Evans ITD, and a set of 51 to the Vaswani ITD: examples of these queries included *AMPLIF*, COMPUT*, *ELASTIC, *METER, TELE*, and *VINYL*, where * represents the presence of an arbitrary substring (which may be empty). The results obtained are shown in Table 3. A very few non-related terms were retrieved using digrams, these corres-

```
Query term  :  *PLANE
Index Terms :  AEROPLANE, AIRPLANE, BIPLANE,
    FOREPLANE, INPLANE, MIDPLANE, MONOPLANE,
    PLANE, TAILPLANE

Query term  :  PHOTO*
Index terms :  PHOTOELASTIC, PHOTOELECTRIC,
    PHOTOGRAPH, PHOTOGRAPHIC, PHOTOGRAPHS,
    PHOTOGRAPHY, PHOTOMULTIPLIER, PHOTORECORDING
    PHOTOTHERMOELASTIC, PHOTOTHERMOPLASTIC

Query term  :  *STRUCT*
Index terms :  CONSTRUCTION, STRUCTURAL,
    STRUCTURE, STRUCTURES
```

FIG. 2. Examples of retrieved index terms using truncated query terms based on trigrams

ponding to cases where all of the digrams were present but either in a different order, as with **CHRONO** and MONOCHROMATIC, or in the correct order but with intervening characters, as with **STATION** and STANDARDIZATION. Such erroneous retrievals could be eliminated before display by an exact match of the

index term against the query term: this would not be very time-consuming in view of the small number of words retrieved. Even without such a matching routine, however, the searches in Figure 2 and the figures in Table 3 show that the suggested method is an effective means of identifying words containing a truncated substring.

5. CONCLUSIONS

The proposed technique can replace a dictionary browsing command, as exemplified by EXPAND or NEIGHBOUR, with the advantages that it can identify related terms to a query word, such as spelling mistakes or variants containing a different prefix, even if they are alphabetically distanced from the query term, and that it enables searches to be made for arbitrarily truncated terms. The results show that the technique can retrieve a high proportion of the words related to a query term and still maintain an acceptable level of precision: thus, using trigrams and the 12000 word Vaswani dictionary, less than 20 words on average were displayed at the terminal even with the lowest threshold similarity tested, these results corresponding to recall and precision figures of 0.88 and 0.53, respectively. Extrapolation to larger files is obviously problematical, but it is felt that dictionaries containing some tens of thousands of words could be searched in this manner without undue difficulty.

ACKNOWLEDGEMENTS

Thanks are due to Dr. P. K. T. Vaswani, Dr. K. Sparck Jones and Mr. L. Evans for data, and to Prof. M. F. Lynch and Miss F. E. Wood for helpful comments.

REFERENCES

Adamson, G. W. and Boreham, J. (1974) The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval 10*, 253–260.

Alberga, C. N. (1967) String similarity and misspellings. *Communications of the ACM 10*, 302–313.

Aldefeld, B., Levinson, S. E. and Szymanski, T. G. (1980) A minimum-distance search technique and its application to automatic directory assistance. *Bell System Technical Journal 59*, 1343–1356.

Blair, C. R. (1960) A program for correcting spelling errors. *Information and Control 3*, 60–67.

Bourne, C. P. (1977) Frequency and impact of spelling errors in bibliographic data bases. *Information Processing and Management 13*, 1–12.

Bourne, C. P. and Ford, D. J. (1961) A study of methods for systematically abbreviating English words and names. *Journal of the ACM 8*, 538–552.

Cercone, N. (1978) Morphological analysis and lexicon design for natural language processing. *Computers and the Humanities 11*, 199–209.

Damerau, F. J. (1964) A technique for computer detection and correction of spelling errors. *Communications of the ACM 7*, 171–176.

Davidson, L. (1962) Retrieval of misspelled names in an airline's passenger record system. *Communications of the ACM 5*, 169–171.

Findler, N. V. and Van Leeuwen, J. (1979) A family of similarity measures between two strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1*, 116–118.

Freund, G. E. (1981) *An investigation into word conflation techniques as an aid to online searching.* University of Sheffield: unpublished M.Sc. dissertation.

Hall, J. L. and Brown, M. J. (1981) *Online Bibliographic Databases—An International Directory.* London: ASLIB.

Hall, P. A. V. and Dowling, G. R. (1980) Approximate string matching. *Computing Surveys 12*, 381–402.

James, E. B. and Partridge, D. P. (1976) Tolerance to inaccuracy in computer programs. *Computer Journal 19*, 207–212.

Lennon, M., Peirce, D. S., Tarry, B. D. and Willett, P. (1981) An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science 3*, 177–183.

Lovins, J. B. (1968) Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics 11*, 22–31.

Lovins, J. B. (1971) Error evaluation for stemming algorithms as clustering algorithms. *Journal of the American Society for Information Science 22*, 28–40.

Lynch, M. F. (1977) Variety generation—a reinterpretation of Shannon's mathematical theory of communication and its implications for information science. *Journal of the American Society for Information Science 28*, 19–25.

Morgan, H. L. (1970) Spelling correction in systems programs. *Communications of the ACM 13*, 90–94.

Noreault, T., Koll, M. and McGill, M. J. (1977) Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science 28*, 333–339.

Paice, C. D. (1977) *Information retrieval and the computer.* London: Macdonald and Jane's.

Peterson, J. L. (1980) Computer programs for detecting and correcting spelling errors. *Communications of the ACM 23*, 676–687.

Pollock, J. J. (1977) Automatic and manual abbreviation. *Proceedings of the ASIS Annual Meeting 14*, microfiche frames E10–F13.

Raben, J. and Lieberman, D. V. (1976) Text comparison: principles and a program. *The computer in literary and linguistic studies.* (A. Jones and R. F. Churchouse, eds.) pp. 297–308. Cardiff: University of Wales Press.

Sparck Jones, K. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*, 11–21.

Stalker, G. H. (1978) Some notions of 'similarity' among lines of text. *Computers and the Humanities 11*, 199–209.

Toussaint, G. T. (1978) The use of context in pattern recognition. *Pattern Recognition 10*, 189–204.

Ullman, J. R. (1977) A binary *n*-gram technique for automatic correction of substitution, deletion, insertion, and reversal errors in words. *Computer Journal 20*, 141–147.

Willett, P. (1979) Document retrieval experiments using indexing vocabularies of varying size. II. Hashing, truncation, digram and trigram encoding of index terms. *Journal of Documentation 35*, 296–305.

Willett, P. (1981) A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing and Management 17*, 53–60.

Zamora, A. (1980) Automatic detection and correction of spelling errors in a large data base. *Journal of the American Society for Information Science 31*, 51–57.