IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Contribution of Word length in Substitution Error Pattern analysis of Punjabi Typed Text

Meenu Bhagat

Department of Computer Science and Engineering, Punjab University SSG Regional Centre, Hoshiarpur, India

**Corresponding Author: meenubhagat@yahoo.com*

Available online at: www.ijcseonline.org

Accepted: 18/May/2018, Published: 31/May/2018

***Abstract*—** Spelling error pattern analysis of a language is useful in language related technology, such as creation of Natural Language Interfaces, Machine Translation, Optical Character Recognition, Spell Checker and Corrector etc. It includes analysis of various types of errors (insertion, deletion, transposition, substitution, run-on, split word error) Positional analysis, Word length effects, Phonetic errors, First position error analysis, Keyboard effects etc. This paper mainly focuses on the effect of word length in substitution error pattern analysis of Punjabi by doing Statistical Error analysis of Punjabi typed text. It also presents a brief overview of effect of word length on non-word error analysis in Punjabi Typed Text. This paper is based on the analysis done on 20000 misspelled words generated by typists.

***Keywords*—** Addak, Gurmukhi, Non-word, Bindi

I. INTRODUCTION

Word Length(i.e. number of characters) plays an important role in non-word error distribution of typed text .It plays an important role in Natural Language Interfaces, spellchecker, OCR and language related technology development etc .Though considerable work has been done in the area for English and related languages, the Indian Language scenario is still far behind. This paper focuses on the contribution of word length in substitution error pattern analysis of Punjabi typed text that can be further useful in automatic text error correction in Punjabi language, the world's most widely spoken language, giving a statistical report about the distribution of various type of errors (substitution, insertion, deletion, transposition etc.) in Punjabi language. Damerau[1] worked on a technique for computer detection and correction of spelling errors in English language. Pollock and Zamora [2] focused on discovering probabilistic tendencies, such as which letters and position within a word are most frequently involved in errors, with the intent of devising a similarity key based techniques. Kukich[3] has discussed the various techniques for automatically detection and correction of misspellings and the various factors affecting the spelling errors patterns of words in English. Church and Gale[4] have done a Probability scoring for spelling correction. Chaudhuri and Kundu[5] have done a detailed analysis on error pattern generated by Bangla text patterns and made a reversed word dictionary and phonetically similar word grouping based spellchecker for Bangla text. Morris and Cherry [6] devised an alternative technique for using trigram frequency statistics to detect errors.Yannakoudakis and Fawthrop [7-8] sought a general characterization of misspelling behavior. Wagner [9] was the first one to introduce the notion of applying dynamic programming techniques to the spelling correction problem to increase computational efficiency.

II. INTRODUCTION OF GURMUKHI SCRIPT[10]

The word 'Gurmukhi' literally means from the mouth of the Guru. Gurmukhi script is used primarily for the Punjabi language, which is world's most widely spoken language. Gurmukhi script-consists of 41 consonants called *vianjans*, 9 vowel symbols called *laga* or *matras*, 2 symbols for nasal sounds, one symbol for reduplication of sound of any consonant and three half characters.

Consonants				
a	A	e		
		s	h	
k	K	g	G	|
c	C	j	J	\
t	T	f	F	x
q	Q	d	D	n
p	P	b	B	m
X	r	l	v	V
S	^	Z	z	&	L
Vowels	w ,i , I , u , U , y , Y, o , O			
Semi-Vowels	N , ° , `			
Half Characters	*HH*	*R*	*Ĺ*	

Table 1: Gurmukhi Vocabulary

Vowel Consonants: The consonants of first row (a,A,e) are classified as open syllabics and called vowel consonants or semi consonants or "Matra Vahak" due to their inherent property that they are never used in work without any 'Laga' or 'Vowel'.

Root Consonant: The next two consonants are classified as root class consonants.

There are five such categories namely the Kavarg toli, Chavarg toli, Tavarg toli and the Pavarg toli depending upon the different organs like throat, palate, mouth, tongue and lips, using which they are pronounced or from where they originate.

Antim Group: The last but one group consisting of 5 independent consonants (X,r,l,v,V) is called the "Antim" group

Naveen Group: "Naveen" group (S,^,Z,z,&,L) is the last group which has been introduced to accommodate the words of Persian, Arabic and Sanskrit. Punjabi has three diacritics namely bindi (ਅਂ), tippi (ਅੰ) and addak (ਅੱ) used with vowels. These diacritics quite important as their use change the meaning of the words.

III. DATA COLLECTION AND ANALYSIS

Statistical data for the results was collected from Typing Colleges, Professional typists and Government institutions and private printing presses and every document was carefully scrutinized and the misspelled words were manually collected and analyzed. Out of Text containing more than eight lakh words around 20000 misspellings were found. Different type of analysis has been performed on the corpus like word length analysis, Special character analysis First position error analysis etc. Though considerable work has been done on automatic spell checking and correction in English language, for Indian language error correction, it has shown more difficulties than that of English because of Indian Language characteristics. The key reasons for difficulties in automatic text error correction in Punjabi are listed below:

(1) Multiple ways of writing the same word.
(2) Difference between Phonetic utterance and the spelling of that word.
(3) Naveen group elements related problems.
(4) Words borrowed from other languages (English etc).
(5) Phonetically Similar Characters.

IV. STATISTICAL ANALYSIS OF RESULTS

It has been found that about 63% of the errors (SE, DE, IE, TE, SWE, ROE) occur in word length 2,3,4,5. Out of total of 21.70% of four character misspellings ,11.54% (fig 1) errors are due to substitution errors, and out of total 20.33% of five character misspellings, 8.87% errors are due to substitution errors. It has been also found that from the misspelling of word length 2, 3,4,5,6 about 36% of errors are due to substitution errors.

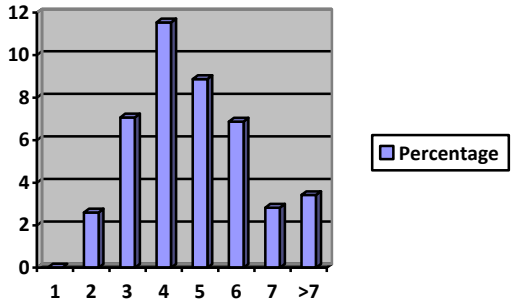

Fig 1: Distribution of substitution error misspellings according to word length

Kukich [3] analyzed over 2000 error types in a English corpus of TDIL conversations. He found that over 63% of the errors occurred in words of length 2, 3, 4 characters. In Punjabi language we analysed that the maximum of the misspellings have word length of five (Fig 2). Out of total 16.19% of six character misspellings, 6.87% (fig 3) errors are due to substitution errors. Out of total 8.73% of seven character misspellings, 2.82% (fig 3) errors are due to substitution errors. It has been observed that about 56% of errors are in words of length 3, 4, and 5. This means words having word length of five contain maximum of errors.

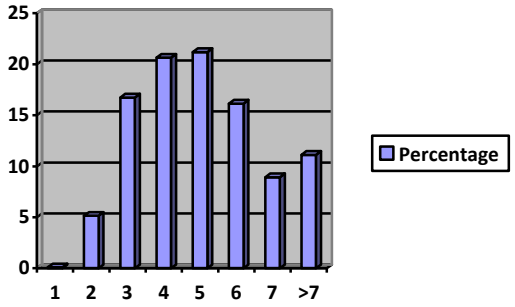

Fig 2: Word Length wise distribution of misspellings

Out of total 16.19% of six character misspellings, 6.87%(fig 3) errors are due to substitution errors.43.20% of errors belongs to substitution type of errors in all types of errors(SE,IE,DE,TE,RE,SWE etc.).

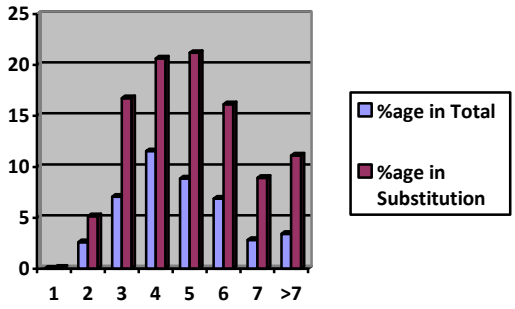

Fig 3: Comparative graph of effect of word length in total no of errors and in substitution errors.

V. CONCLUSION

A detailed study has been made on the different type of errors analysis of Punjabi Typed text regarding automatic text error correction in Punjabi. This analysis is helpful in creating suggestion list for Punjabi spellchecker. I have done analysis based on, positional effects, first position error analysis, phonetic effects, word length effects etc. This paper mainly focuses on effect of Word length in Substitution Error Pattern analysis of Punjabi Typed Text.

Following are the major findings concluded after analysis:

1) It has been observed that 63% of the errors occurred in words of length 2, 3, 4 characters.

2) It has been observed that about 56% of errors are in words of length 3, 4, and 5.

3) It is observed that out of total 21.70% of four character misspellings, 11.54% errors are due to substitution errors

REFERENCES

[1] F.J. Damerau (1964) "A technique for computer detection and correction of spelling errors".*Commun. ACM*. 7(3): 171-176.

[2] POLLOCK, J. J., AND ZAMORA, A. 1983. Collection and Characterization of spelling errors in scientific and scholarly text. J. Amer. Soc. Inf. Sci. 34, 1, 51-58.

[3] K. Kukich (1992) "Techniques for automatically correcting words in text". *ACM Computing Surveys.* 24(4): 377-439.

[4] K.W. Church and W.A. Gale (1991) "Probability scoring for spelling correction". Statistical Computing. 1(1): 93-103.

[5] P. Kundu and B.B. Chaudhuri (1999) "Error Pattern in Bangla Text". *International Journal of Dravidian Linguistics*. 28(2): 49-88.

[6] Morris, Robert & Cherry, Lorinda L, 'Computer detection of typographical errors', *IEEE Trans Professional Communication*, vol. PC-18, no.1, pp54-64, March 1975.

[7] Yannakoudakis, E.J. & Fawthrop, D 1983a. An Intelligent spelling corrector. Inf. Process. Manage. 19, 12, 101-108.

[8] Yannakoudakis, E.J. & Fawthrop, D, 'An intelligent spelling error corrector', *Information Processing and Management*, vol.19, no.2, pp101-108, 1983. (1983b)

[9] Wagner, Robert A. & Fischer, Michael J, 'The string-to-string correction problem', *Journal of the A.C.M.*, vol.21, no.1, pp168-173, January 1974.

[10] Meenu Bhagat, "Contribution of 'Addak'and 'Bindi' in Non word Error Pattern analysis of Punjabi Typed Text", "International Journal of Computer Sciences and engineering" vol. 5 issue 9.

