Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.182 – 186

Contribution of Word Length in Non-word Error Distribution of Punjabi Typed Text

Meenu Bhagat

Department of Computer Science & Engg., Punjab University SSG Regional Centre Hoshiarpur, Punjab, India

meenubhagat@yahoo.com

Abstract— 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 related errors in Non-word Error distribution 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 This paper is based on the analysis done on 20000 misspelled words generated by typists.This paper also compares the word length effect of Punjabi language with English language.

Keywords— Word length, Phonetic, Kavarg, Naveen, Gurmukhi

I. INTRODUCTION

Kukich[1] 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. Chaudhuri and Kundu[2]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. Church and Gale[3] have done a Probability scoring for spelling correction. Damerau[4] worked on a technique for computer detection and correction of spelling errors in English language. Morris and Cherry [5] devised an alternative technique for using trigram frequency statistics to detect errors. Pollock and Zamora [6] aimed at 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 technique. 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.

A "reverse" minimum edit distance technique was used by Gorin [10] in the DEC-10 spelling corrector and by Durham et al.[11] in their command language corrector. Kernighan et al [12] and Church and Gale [13] also used a reverse technique to generate candidates for their probabilistic spelling corrector.

II. A BRIEF OVERVIEW OF GURMUKHI SCRIPT (14)

The word 'Gurmukhi' literally means from the mouth of the Guru. Gurmukhi script is used primarily for the Punjabi language, which is world's 14th most widely spoken language. Punjabi is named after Punjab, which was divided between India and Pakistan during Partition in 1947. Punjab literally means land of five rivers; Punj meaning five and Aab, water. Gurmukhi script is syllabic in nature. 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.

Consonant

a	A	e				Matra Vahak
s	h					Mul Varag
k	K	g	G	|		Kavarg Toli
c	C	j	J	\		Chavarg Toli
t	T	f	F	x		T ææavarg Toli
q	Q	d	D	n		Tavarg Toli
p	P	b	B	m		Pavarg Toli
X	r	l	v	V		Antim Toli
S	^	Z	z	&	L	Naveen Toli

Vowels

w ,i , I , u , U , y , Y, o , O

Semi-Vowels

N , ° , `

Half Characters

HH R Í

Table 1: Gurmukhi Vocabulary

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'. The next two consonants are classified as root class consonants. The rest of the consonants except to the last two groups namely the - "Antim" and "Naveen" group, are categorized according to their phonetic structure.

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.

The last but one group consisting of 5 independent consonants (X,r,l,v,V) is called the "Antim" group and the last group is the (S,^,Z,z,&,L). "Naveen" group which has been introduced to accommodate the words of Persian, Arabic and Sanskrit.

III. Data Collection and Analysis

Material was collected from Type Colleges, Professional typists and Government institutions and private printing presses and every document was carefully checked and the misspelled words were manually collected and analyzed. Out of Text containing more than eight lakh words around 20000 misspellings were found.

IV. STATISTICAL ANALYSIS OF RESULTS

In English **Kukich[1]** analyzed over 2000 error types in a corpus of TDIL conversations and found that over 63% of the errors occurred in words of length 2,3,4 characters. According to our results the maximum of the misspellings have word length of five(Fig 1). It is 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.

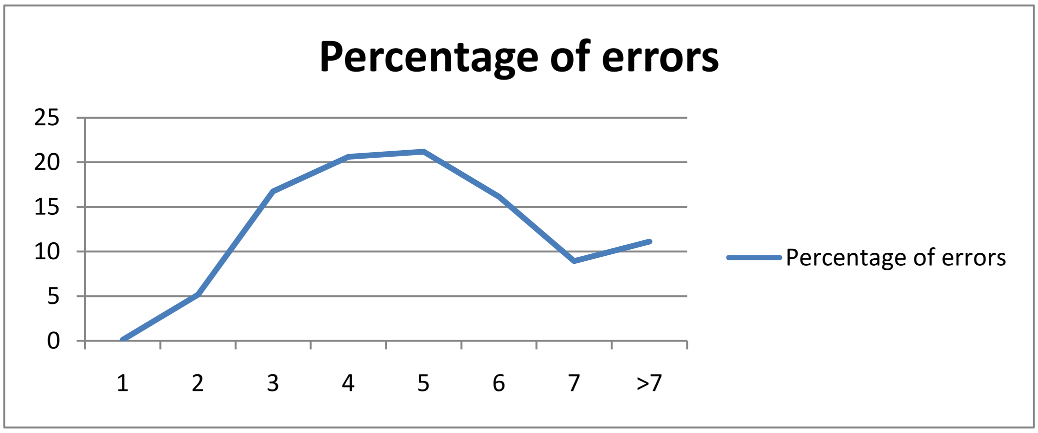

Fig 1: Word Length wise distribution of misspellings

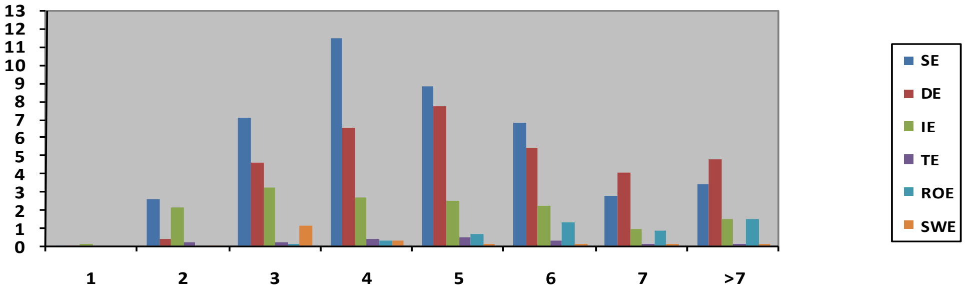

Fig 2 Percentage of various types of errors in various word length zones

is showing the Percentages of various types of errors in various word length zones. It is seen that about 63% of the errors (SE, DE, IE, TE, SWE, ROE) occur in word length 2,3,4,5.Out of total 21.30% of four character misspellings, 11.54% errors are due to substitution errors and similarly out of total 20.33 % age of five character misspellings, 8.87% errors are due to substitution errors. In the misspelling of word length 2, 3,4,5,6 about 36% of errors are substitution errors.

V. CONCLUSION

A detailed study has been made on effect of word length in Non word error distribution of Punjabi Typed text for the automatic text error correction in Punjabi. The results of this analysis are 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. Following points have been concluded regarding the contribution of word length in overall error distribution:

1. 32.91% of the first position misspellings are due to substitution errors.
2. 63% of the errors (SE, DE, IE, TE, SWE, ROE) occur in word length 2,3,4,5
3. 11.54% errors are due to substitution errors.
4. Due to phonetic similarities of various consonants and vowels.

REFERENCES

[1] K. Kukich (1992) "Techniques for Automatically Correcting words in Text". ACM Computing Surveys. 24(4): 377-439.

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

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

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

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

[6] 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.

[7] YANNAKOUDAKIS, E. J., AND 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] R.E. Gorin (1971) "SPELL: A spelling checking and correction program", *Online documentation for the DEC-10 computer.*

[11] Durham, I, Lamb, D.A, & Saxe, J.B, 'Spelling correction in user interfaces', *Communications of the A.C.M.*, vol.26, no.10, pp764-773, October 1983.

[12] M.D. Kernighan, K.W. Church, and W.A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, pages 205-210.

[13] Gale and Church, 1991[b] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Meeting of the ACL,* pages 177-184. Association for Computational Linguistics, 1991.

[14] Meenu Bhagat,"Difficulties in automatic text error correction in Punjabi", International Conference on Control Communication and Computer Technology" 6-7th Aug, New Delhi.