ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

A Statistical Review of Word length effect in Non word Error distribution of Punjabi Typed Text and English Language

Meenu Bhagat

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

ABSTRACT: Word Error can be of two types: Non word and Real word.Though considerable work has been done in the area for English and related languages regarding types of errors and their characteristics, the Indian Language scenario is still far behind. This paper gives a comparison of word length effect of Non word error distribution in Punjabi Typed Text and English language. This paper is based on the analysis done on 20000 misspelled words generated by typists. This type of analysis can be useful in Natural Language Interfaces, spellchecker, OCR and language related technology development etc.

KEYWORDS*:* Word length, Phonetic, Kavarg, Naveen, Gurmukhi.

I. INTRODUCTION

Error can be of two types, namely, Non-word error and Real-word error. If a string of characters is separated by spaces or punctuation marks it is called a Candidate string. A Candidate string is said to be valid word if it has a meaning. Else, it is a nonword.Real Word error is a valid word but not the intended word in the sentence, making the sentence syntactically or semantically incorrect. In each case the problem is to detect the Error and suggest correct alternatives or automatically replace it with correct word.

Kukich[1] has discussed the different techniques for automatically detection and correction of misspellings and the identify the various factors affecting the spelling errors patterns of words in English. Damerau [2] worked on a technique for computer detection and correction of spelling errors in English language. Church and Gale [3] have done a probability scoring for spelling correction. Chaudhuri and Kundu [4] have done an elaborative analysis on error pattern generated by Bangla text patterns and made a reversed word dictionary and phonetically similar word grouping based Bangla spellchecker. Pollock and Zamora [5] 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. 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 behaviour. Wagner [9] was the first to introduce the concept 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. Church and Gale [12] and Kernighan et al [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.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

The consonants of first row (a,A,e) are classified as open syllabics and called vowel consonants or semi consonants or "MatraVahak" 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 Kavargtoli, Chavargtoli, Tavargtoli and the Pavargtoli 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. STATISTICAL ANALYSIS OF RESULTS

According to Zipfs law [15], short words occur more frequently than long words. As per study conducted by Landauer and Streeter [16], high-frequency (i.e., short) words tend to have more single-error neighbors than low-frequency words, thus making it difficult to select the intended correction from its set of neighbors. Pollock and Zamora's [5], studied that of 50,000 non-word errors indicated that errors in short words are indeed problematic for spelling correction even if their frequency of occurrence may be low. They state: "although 3–4 character misspellings constitute only 9.2% of total misspellings, they generate 42% of the miscorrections" .For example, Yannakoudakis and Fawthrop [17] found an even lower frequency of occurrence of errors in short words, about 1.5%, in their analysis of 1,377 errors found in the literature. Kukich [1] analyzed over 2,000 error types in a corpus of TDD conversations and found that over 63% of the errors occurred in words of length 2, 3, and 4 characters. These differences emphasize the need to know the actual characteristics of the spelling errors for any application well before designing or implementing a correction system.

In Punjabi Language [18], It has been seen that the maximum of the misspellings have word length of five. It is observed that about 56% of errors are in words of length 3, 4, and 5. (Fig 1).
This means words having word length of five contain maximum of errors.

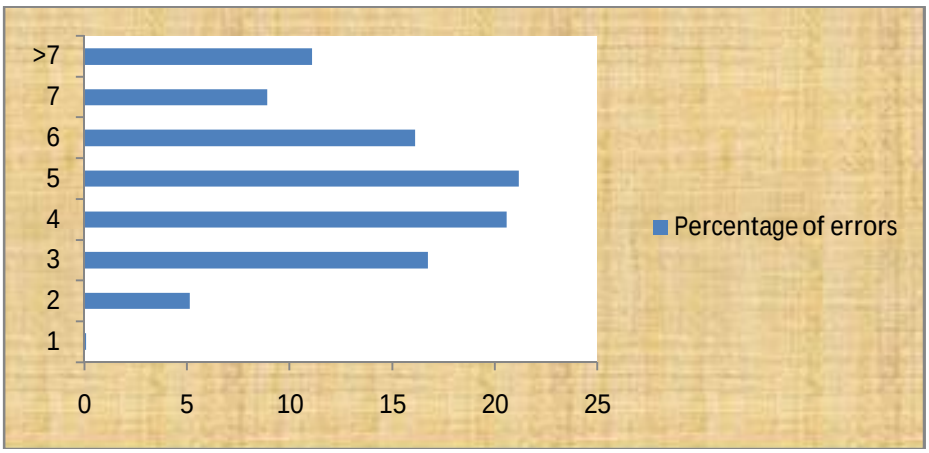

Fig 1: Word Length wise distribution of misspellings

Fig 2 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 % of five character

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

misspellings, 8.87% errors are due to substitution errors.

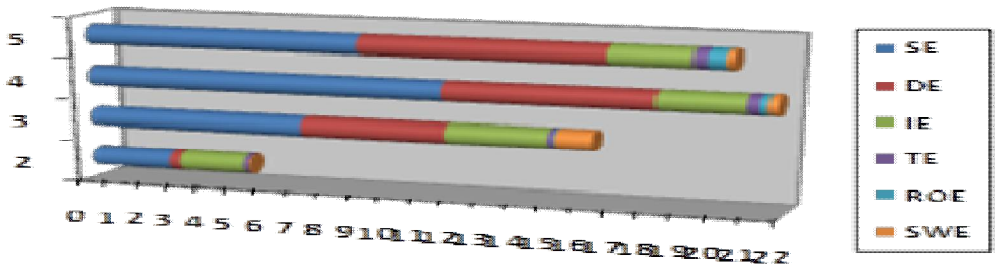

Fig 2 Percentage of various types of errors in various word-length zones

In the misspelling of word length 2, 3,4,5,6 about 36% of errors are substitution errors.

IV. CONCLUSION

A Comparative study has been made on effect of word length in Non word error distribution of Punjabi and English language. 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:

a) Kukich [1990] analyzed over 2,000 error types in a corpus of TDD conversations and found that over 63% of the errors occurred in words of length 2, 3, and 4 characters. In Punjabi language as per our results 56% of errors are in words of length 3, 4, and 5.
b) 63% of the errors (SE, DE, IE, TE, SWE, ROE) occur in word length 2,3,4,5
c) Out of total 20.33 % age of five character misspellings, 8.87% errors are due to substitution errors.
d) 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] F.J. Damerau (1964) "A Technique for computer detection and correction of spelling errors".*Commun. ACM.* 7(3): 171-176.
[3] K.W. Church and W.A. Gale (1991) "Probability scoring for Spelling correction". Statistical Computing. 1(1): 93-103.
[4] P. Kundu and B.B. Chaudhuri (1999) "Error Pattern in Bangla Text". *International Journal of Dravidian Linguistics*. 28(2): 49-88.
[5] 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.
[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., 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.*

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

[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] 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.
[13] 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.
[14] MeenuBhagat,"Difficulties in automatic text error correction in Punjabi", International Conference on Control Communication and Computer Technology" 6-7th Aug, New Delhi.
[15] ZIPF, G. K. 1935. The Psycho-Bzology of Language. Houghton Mifflin, Boston.
[16] LANDAUER, T. K, AND STREETER. L. A. 1973. Structural differences between common and rare words,
[17]YANNAKOUDAKIS, E. J , AND FAWTHROP, D. 1983b.The rules of spelling errors. Znf. Process. Manage.19, 2, 87–99.YOUNG,
[18] MeenuBhagat, (2007), "Spelling Error Pattern Analysis of Punjabi Typed Text", Thesis report, Thapar University, Patiala.