ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 7, July 2016

A Brief Review of Single/Multi-Error Misspellings in English, Spanish, Punjabi and Arabic

Meenu Bhagat
Assistant Professor,
Department of Computer Science & Engineering
Punjab University SSG Regional Centre Hoshiarpur, Punjab, India
meenubhagat@yahoo.com

Abstract: **Single error misspellings are the misspellings in which a word contains one error where as multi-error misspellings where a single word contains multiple instances of errors. This paper focuses on the contribution of Single/Multi-Error misspellings in Punjabi Typed Text and compares it with other languages like Spanish, English and Arabic. The statistical data we provide on spelling error patterns in Punjabi and their comparison with other data in other languages are the novel contribution of this paper.**

Keywords**: Kavarg, Naveen, Gurmukhi, Non-word.**

I. INTRODUCTION

Error can be of two types, 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 otherwise, it is a non-word. Real Word error is a valid word but not the intended word in the sentence, making the sentence syntactically or semantically incorrect.

Damerau [1] worked on a technique for computer detection and correction of spelling errors in English language. Kukich[2] has discussed the different techniques for automatically detection and correction of misspellings and identification of the various factors affecting the spelling errors patterns of words in English. Church and Gale [3] have done a probability scoring of 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.

In each case the problem is to detect the Error and suggest correct alternatives or automatically replace it with correct word.

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 æavarg 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

Last group is the (S,^,Z,z,&,L). "Naveen" group which has been introduced to accommodate the words of Persian, Arabic and Sanskrit.

II. SINGLE/MULTI-ERROR DISTRIBUTION IN ENGLISH

Single error misspellings are the misspellings in which a word contains one error .In multi-error misspellings a single word contains multiple instances of errors. Kukich [1] has found upwards of 80% misspellings to be single error misspellings and most misspellings tend to be within two characters in length of the correct misspelling. Damerau[2] found that approximately 80% of all misspelled words contained a single instance of one of the following four types of errors: **insertion**, **deletion**, **substitution** and **transposition**.An analysis was also carried for different type of Single/Multi-error misspellings for Punjabi typed Text and It has been found that out of the total no. of misspellings, 91.13% were the single error misspellings and 8.87% were multi error misspellings. While for English language, **Pollock and Zamora[5]** found that only 6% of 50000 nonword spelling errors in the machine readable databases they studied were multierror misspellings and Coversely , **Mitton[15]** found that 31% of the misspellings in his 17001 word corpus of handwritten essays contained multiple errors.

III. SINGLE/MULTI-ERROR MISSPELLINGS IN SPANISH LANGUAGE (16)

In Spanish Language, Ramirez Bustamante and F., E. López Diaz [16] found that vast majority of errors found in the corpus are single error misspellings (over 89%).Multi-error misspellings are less than 9%. There is an insignificant remaining percentage of noise related to spaces in multiple locations, extreme multi-error words and indecipherable strings of characters. The corpus used contains 8 million words of edited and unedited texts.

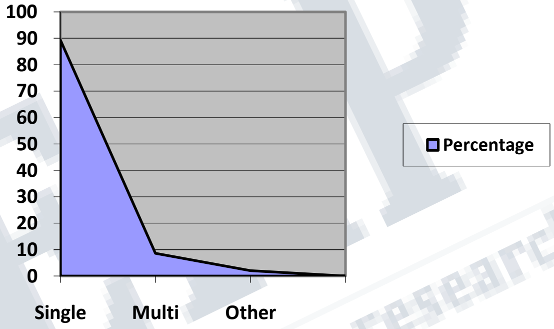

Fig1 Showing Percentage of single and multi-error misspellings

IV. SINGLE/MULTI-ERROR MISSPELLINGS IN PUNJABI LANGUAGE

An analysis was also carried for different type of Single/Multi-error misspellings for Punjabi typed Text and It has been found that out of the total no. of misspellings, 91.13% were the single error misspellings and 8.87% were multi error misspellings.

It is observed that majority of the multi-error misspellings contain two mistakes (see Fig 2).

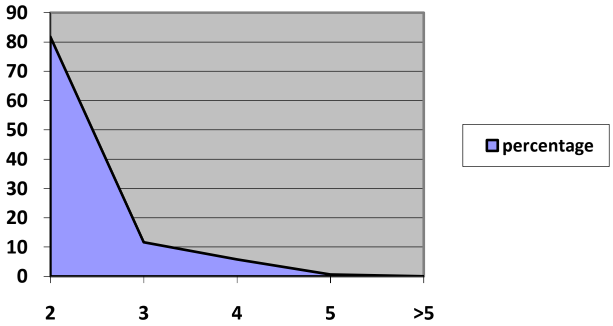

Fig 2 Showing the Percentages of no. of mistakes in a word

The positional analysis plays an important and significant factor in the error pattern study. This can lead us to error zone of high probability. It has been found out that patterns for the positional mistakes are almost similar in both single/multi-error misspellings. The maximum of the mistakes occur at the third position and the error zone decreases after 3rd position.

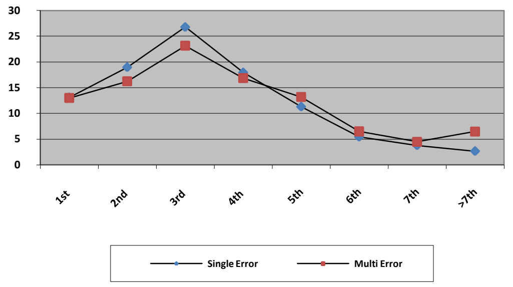

Figure 3 Position wise distribution of Single/Multi-error misspellings

It is observed that in single error misspellings 13.10% and 13.0% in multi error misspellings are found to be first position errors.

V. SINGLE/MULTI-ERROR MISSPELLINGS IN ARABIC LANGUAGE[17]

Based on the analysis presented in [18] 80% of all misspelling errors in Arabic refer to single error misspellings

VI. CONCLUSION

Spelling error pattern analysis of a language is helpful in language related technology, such as creation of Spell Checker and Corrector, Optical Character Recognition, Machine Translation, Natural Language Interfaces etc. It includes analysis of various types of errors transposition, substitution, insertion, deletion, run-on, split word error phonetic errors, keyboard effects etc.This paper reports findings from previous generalizations about spelling error patterns found in other languages like Arabic, Spanish English and Punjabi and offers new insights on them. This analysis presents a brief overview of single/multi-error misspellings in different languages done on different types and different amount of text.
Following points are concluded:

1 In Punjabi out of the total no. of misspellings, 91.13% were the single error misspellings and 8.87% were multi error misspellings.

2 In English, Kukich [1] has found upwards of 80% misspellings to be single error misspellings and most misspellings tend to be within two characters in length of the correct misspelling.

3 In Spanish Language, Ramirez Bustamante and F., E. López Diaz [16] found that over 89% are single error misspellings .Multi-error misspellings are less than 9%.

4 In Arabic [18] 80% of all misspelling errors in Arabic refer to single error misspellings

REFERENCES

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

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

[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.

[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] Meenu Bhagat,"Difficulties in Automatic Text Error Correction in Punjabi", International Conference on Control Communication and Computer Technology" 6-7th Aug, New Delhi.

[15] Roger Mitton (1987),"Spelling checkers, spelling correctors and the misspellings of poor spellers", Information Processing and Management: an International Journal, v.23, and pp. 495-505.

[16]Ramirez Bustamante, F., E. López Díaz. Spelling Error Patterns in Spanish for Word Processing Applications. Proceedings of LREC 2006 (http://pages.cs.brandeis.edu/~marc/misc/proceedings/lrec-2006/pdf/119_pdf.pdf).

[17]Bassam Haddad, Mustafa Yaseen,"Detection and Correction of non words in Arabic: A Hybrid Approach", International Journal of Computer Processing of Oriental Languages Volume 20, Number 4, 2007.

[18]Ben Hamadou ,A., "The Phases of Computational Analysis of Arabic towards detecting and correcting of errors,"Second Conference for Arabization of Computers in Arabic,1994.