Error pattern in Punjabi Typed Text

G S Lehal[1] and Meenu Bhagat[2]

[1] Department of Computer Science & Engineering, Punjabi University, Patiala, India.

gslehal@mailcity.com

[2] Department of Computer Science, DAV Institute of Engineering & Technology, Jalandhar, India.

Abstract

Error pattern analysis of a language is useful in language related technology development, such as Spell Checker and Corrector, Optical Character Recognition, Machine Translation, Natural Language Interfaces etc. Error pattern analysis 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. Though considerable work has been done in the area for English and related languages, the Indian Language scenario presents a relatively more complex and uphill task. In this paper, we have presented a statistical error analysis for Punjabi, the world's 14th most widely spoken language. For this purpose we have collected about 20000 misspelled words generated by typists.

1 Introduction

Error pattern analysis of a language is useful in language related technology development, such as Spell Checker and Corrector, Optical Character Recognition, Machine Translation, Natural Language Interfaces etc. Error pattern analysis 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.

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.

This is the first time that a detailed error analysis for Punjabi is being carried out. For this purpose we have collected about 20000 misspelled words generated by typists, both novice and experienced as well as students learning Punjabi typing. We have done analysis of six main categories of errors. These errors are discussed in detail in following sections.

2 A Brief Overview of Gurmukhi Script

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. 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.

Table 1: Gurmukhi Vocabulary

Consonant

ੳ	ਅ	ੲ			Matra Vahak
			ਸ	ਹ	Mul Varag
ਕ	ਖ	ਗ	ਘ	ਙ	Kavarg Toli
ਚ	ਛ	ਜ	ਝ	ਞ	Chavarg Toli
ਟ	ਠ	ਡ	ਢ	ਣ	Ṭ avarg Toli

ਤ	ਥ	ਦ	ਧ	ਨ		Tavarg Toli
ਪ	ਫ	ਬ	ਭ	ਮ		Pavarg Toli
ਯ	ਰ	ਲ	ਵ	ੜ		Antim Toli
ਸ਼	ਖ਼	ਗ਼	ਜ਼	ਫ਼	ਲ਼	Naveen Toli

Vowels

ਾ ,ਿ , ੀ , ੁ , ੂ , ੇ , ੈ, ੋ , ੌ

Semi-Vowels

ਂ , ੰ , ੱ

Half Characters

੍ਹ ੍ਰ ੍ਵ

The first three consonants are 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 three consonants (ੳ,ਅ,ੲ) 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 (ਯ,ਰ,ਲ,ਵ,ੜ) is called the "Antim" group and the last group is the (ਸ਼,ਖ਼,ਗ਼,ਜ਼,ਫ਼,ਲ਼). "Naveen" group which has been introduced to accommodate the words of Persian, Arabic and Sanskrit.

3 Difficulties in Automatic Text Error Correction in Punjabi

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

In Punjabi there are many ways of writing the same word and all the ways could be correct. In Punjabi language there is no standardization of spellings and for majority of common words multiple spellings are used and the same word has been written in different forms with different set of characters.e.g. ਪਰਿਚੈ →ਪਰੀਚੈ→ਪਰਿਚਯ →ਪਰੀਚਯ. All the four words are delivering the same meaning and are the different iterations of the same word. So it is affecting the estimated dictionary size and structure.

The common causes for multiple spellings are:

- Usage of the characters of naveen group e.g. ਆਲੇ-ਦੁਆਲੇ → ਆਲ਼ੇ-ਦੁਆਲ਼ੇ
- Usage of ੱ e.g. ਸਾਹਿਤ → ਸਾਹਿੱਤ
- Usage of ਿ e.g. ਯੋਗਤਾ → ਯੋਗਿਤਾ
- Usage of half characters ਕ੍ਰਮ →ਕਰਮ
- Confusion between ਿ , ੀ e.g. ਪਰਿਚੈ →ਪਰੀਚੈ→ਪਰਿਚਯ →ਪਰੀਚਯ

(2) Difference between phonetic utterance and the spelling of that word

There are many words for whose pronunciation is different from its spelling e.g. ਸ਼ੈਹਰ→ਸ਼ਹਿਰ,ਔਂਦਾ→ਆਉਂਦਾ,ਆਯਾ→ਆਇਆ where ਸ਼ੈਹਰ,ਔਂਦਾ,ਆਯਾ are the pronunciations of the words ਸ਼ਹਿਰ,ਆਉਂਦਾ,ਆਇਆ respectively.

(3) Problems regarding the characters of Naveen group

Problems are seen due to the characters of Naveen group i.e. (ਸ਼→ਸ ,ਖ਼→ਖ,ਗ਼→ਗ,ਜ਼→ਜ , ਫ਼→ਫ,ਲ਼→ਲ). These characters only differ by the presence or absence of nukta symbol and in many cases they are pronounced almost similarly. It is seen that maximum of the users confuse during typing the exact character.

It is analysed that the percentage of occurrence of these characters is in the order of ਸ਼ > ਜ਼ > ਖ਼ > ਫ਼ > ਲ਼ > ਗ਼ and the order of mistyping of a word containing these characters is ਸ਼ < ਜ਼ < ਲ਼ < ਖ਼ < ਫ਼ < ਗ਼ .

(4) Borrowed words from other Languages

Modern Punjabi language has many words borrowed and/or assimilated from other languages especially from English, Urdu and Hindi. Sometimes typist knows the word in other language but he makes mistake during typing that foreign word in Punjabi. And it is seen that maximum of the borrowed words are misspelled in different forms e.g. ਟਰੈਕਟਰ→ ਟ੍ਰੈਕਟਰ → ਟਰੈੱਕਟਰ ,ਕਿਰਪਾ→ ਕ੍ਰਿਪਾ and ਅਸਤਿਤਵ→ਅਸਤਿਤ੍ਵ→ਅਸਤਿੱਤਵ etc.

4 Data Collection and Analysis

We have collected the material 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.

As we have discussed earlier about multiple forms of a word, so it might be wrong to collect the raw typed text as the data for analysis. Because analysis of that raw text doest not surely direct us to the *typing mistake* but can mislead us to the *spelling mistake* of that word. e.g. ਸਭਿਆਚਾਰਕ, ਸੱਭਿਆਚਾਰਕ ,ਸੱਭਿਆਚਾਰਿਕ are the different iterations of the same word and different linguists or Punjabi dictionaries are using all of the above forms according to their knowledge. Our main interest is to analyse the typing mistakes instead of spelling mistakes since the study will be used to design a suggestion list for a Punjabi Spellchecker. We have made a careful analysis of each and every word and collected information like single/multi-error misspellings, mistake positions and word length analysis, types of errors for single/multi-error misspellings, special character errors, errors related to vowels, phonetic occurrences etc.

5 Nature of Errors

Damerau[4] 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**. In addition to these another category of errors known as run-on and split word error is also commonly found.Misspellings that fall into this large class are often reffered to as **single error misspellings**; misspellings that contain more than one such error have been dubbed **multi-error misspellings.**

We have divided our analysis work into following categories:

1. Type of error: Substitution, Insertion, Deletion, Transposition, Run-on, Split word error.
2. Positional analysis: Based on the position at which mistake occurs.
3. Word length Effect: Analysis based on the number of characters in the word.
4. Number of mistakes in a misspelling
5. Phonetically Similar Character Analysis
6. First Position error Analysis
7. Other Findings

All the collected misspellings were sorted out for single/multi-error misspellings. 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 [1984][6]** found that only 6% of 50000 nonword spelling errors in the machine readable databases they studied were multierror misspellingsand Coversely , **Mitton [1987][14]** found that 31% of the misspellings in his 17001 word corpus of handwritten essays contained multiple errors.

5.1 Error Pattern based on Type of Error

We have done an analysis of six types of errors i.e.

i. **Insertion error (IE):** When at least one extra character is inserted in the desired word.

ii. **Deletion error (DE):** When at least one character is deleted in the desired word.

iii. **Substitution error (SE):** When at least one character is substituted by the other character. The maximum of misspellings in Punjabi contain substitution errors.

iv. **Transposition error (TE):** When two adjacent characters are transposed.

v. **Run-on Error (ROE):** When there is space missing between two or more valid words**.**

vi. **Split Word error (SWE):** This is Opposite of Run-on error when there is some extra space is inserted between parts of a word. The error can be removed by removing the extra space.

It is analysed that error rate is at its peak due to substitution errors in single as well as multi-error misspellings. In single error misspellings 42.17% and in multi-error misspellings 47.91 % of substitution error rate is found. The reasons for the maximum substitution rate are discussed in the later sections. Table 1 is showing the detailed statistics of the various types of errors in single/multi-error misspellings.

Table 2 Percentages of various types of errors

Type of Error	**SE**	**DE**	**IE**	**TE**	**ROE**	**SWE**
%Age in Single error misspellings	42.17	33.78	14.68	1.85	5.20	2.32
%Age in Multi-error misspellings	47.91	32.84	17.0	1.43	0.60	0.22

While in Bangla[2] for the text containing 1,24,431 misspellings, %ages of Substitution, Deletion, Insertion and Transposition errors are 66.32, 21.88, 6.53 and 5.27 respectively. Thus the substitution and Transposition error rates are high in Bangla as compared to Punjabi. While deletion and insertion error rates are low.

Comparison Of Various Types of Errors in Single/Multi-Error Misspellings

The order of error for various types of errors in single error misspellings is SE> DE> IE> ROE > SWE> TE while in multi-error misspellings is SE> DE> IE> TE> ROE> SWE.Run-on error and split word error are found to be lesser in multi-error misspellings.

5.1.1 Substitution error Analysis

This error occurs when at least one character is substituted by the other character. e.g. ਉਸਦਾ→ਵੁਸਦਾ,ਸ਼ਾਲ →ਸਾਲ, ਸੁਣਦਾ→ ਬੁਣਦਾ etc. In the above three words e.g. ਓ→ਵ,ਸ਼→ਸ,ਸ→ਬ

are the various substitution character pairs respectively. Table 2 is showing the contribution of various substitution character combinations. It can be observed from Table 2 that the top 6 pairs contribute to more than 24% of the substitution errors

The common reasons for substitution errors are:

i. Naveen Group elements: It is seen that 9.91% of the substitution errors are due to the naveen group elements. e.g. ਜ਼ੇਲ→ ਜੇਲ, ਸ਼ਹਿਦ →ਸਹਿਦ.

ii. Due to assignment to same keys (shifted and unshifted modes) on the keyboard: e.g. ਠ →ਲ,ੳ→ਵ , ਣ→ਦ .

iii. Words that are usually used in various forms e.g. ਜੇਹਾ →ਜਿਹਾ, ਕ੍ਰਮ→ ਕਰਮ.

iv. Vowels having similar sounds e.g. ੇ →ੈ, ੋ→ ੌ , ੁ→ ੂ , ਿ → ੀ.

v. Due to substitution of half characters e.g. ਰ→ ੍ਰ ,ਵ → ੍ਵ , ਹ → ੍ਹ .

Table 3: Common wrongly typed character pairs

Sr.no.	Wrongly Typed character pair	%age out of total no. of substitution errors	Cumulative %age	%age out of total no. of errors
1	ਸ਼→ਸ	6.13	6.13	2.65
2	ਫ਼→ਫ	4.51	10.64	1.95
3	ਖ਼→ਖ	3.85	14.49	1.66
4	ਣ→ਦ	3.54	18.03	1.53
5	ਜ਼→ਜ	3.09	21.12	1.34
6	ਠ→ਣ	2.95	24.07	1.27
7	–→=	2.81	26.88	1.21
8	ਰ→ ੍ਰ	2.54	29.42	1.10
9	ੋ → ੌ	2.38	31.80	1.03
10	ਠ →ਲ	2.06	33.86	0.89
11	ਗ਼→ਗ	1.82	35.68	0.79
12	ੇ→ ੈ	1.64	37.32	0.71
13	ੳ→ ਵ	1.34	38.66	0.58
14	ਲ਼→ਲ	1.28	39.94	0.55
15	ੲ →ਹ	1.06	41.00	0.45
16	ਤ →ਦ	0.91	41.91	0.39
17	˘ →º	0.72	42.63	0.31
18	ਵ→ ੍ਵ	0.43	43.06	0.18

Note :→ is showing the bi-directional confusion

5.1.2 Deletion error Analysis

Deletion error: When at least one character is deleted in the desired word e.g. ਗੱਲ →ਗਲ ,ਚੱਲ→ ਚਲ .These errors also give rise to real word errors e.g. in the following example

ਫੁੱਲ→ ਫੱਲ, ਪਾਣੀ →ਪਾਣ etc. ਫੱਲ,ਪਾਣ are two valid words also but they are not the desired word. It is observed that deletion related errors contribute significantly after substitution errors. It is seen that the characters ੱ ,ਂ characters are most commonly missing characters. The percentage of missing ੱ is 20.51 %and the percentage of missing ਂ is 17.47% and these two characters along contribute to 38% of deletion errors. Table 4 shows the % ages of most commonly missing characters.

Table 4 : Commonly missing characters

Sr.no.	**Character**	**%age out of total no. of deletion errors**	**Cumulative %age**
1	ੱ	20.51	20.51
2	ਂ	17.47	37.98
3	ਾ	5.54	43.52
4	ੂ	4.07	47.59
5	—	3.72	51.31
6	ੁ	3.20	54.51
7	ੋ	3.13	57.64
8	ਿ	2.96	60.60
9	ੀ	2.30	62.90
10	੍	0.32	63.22
11	੍ਹ	0.03	63.25

5.1.3 Insertion Error Analysis

Insertion error: When at least one extra character is inserted in the desired word. e.g.

ਜ਼ਹਰ →ਜ਼ਹਿਰ , here ਿ is the extra inserted character. These errors also give rise to real word errors e.g. ਯੋਗਤਾ →ਯੋਗਿਤਾ, ਸਾਰਾ →ਸਾਰ.

In the above example ਯੋਗਿਤਾ,ਸਾਰ are two valid words but they are not the desired word. In the multiple form words confusion regarding insertion errors are due:

1. The use of ੱ e.g. ਸਿਖਿਆ→ਸਿੱਖਿਆ words on both side are delivering the same meaning.
2. The use of ਿ e.g. ਸਾਹਿਤ→ਸਾਹਿੱਤ words on both side are delivering the same meaning.

It is seen that the characters ਂ, ੱ characters are mostly extra inserted characters. The percentage of insertion ਂ is 17.53 % and the percentage of ੱ is 12.52% and these two characters contribute around 30% of Insertion errors. (see Table 5).

Table 5: Most Commonly insertion errors

Sr.no.	Character	%age out of total no. of insertion errors	Cumulative %age
1	ਂ	17.53	17.53
2	ੱ	12.52	30.05
3	ਾ	7.33	37.38
4	ੰ	4.14	41.52
5	ੋ	3.52	45.04
6	ਿ	2.44	47.48
7	ੁ	2.09	49.57
8	ੀ	1.88	51.45
9	—	1.49	52.94

1.1.4 Transposition error analysis

Transposition error occurs when two adjacent characters of the word are typed in swapped manner.e.g. ਸਵੇਰ→ ਸਵਰੇ ,ਰਾਤ→ ਰਤਾ. In the above two words ੇ→ਰ ,ਾ →ਤ are transposed character pairs.

It is found that these transpositions (like substitution) also give rise to real word errors e.g.

ਕਰਮ →ਕਮਰ, ਸੂਰਤ→ਸੂਤਰ where ਕਮਰ,ਸੂਤਰ are two valid words. The %age of transposition errors is 1.85% and 1.43% in single and multi-error misspellings respectively. No prominent transposition character pairs were found.

1.1.5 Run-on errors

Run-on Error[2]: This type of error occurs when two or more valid words are mistakenly written side by side without a space in between[2] . e.g. ਜਿਸ ਦਾ→ ਜਿਸਦਾ,ਦਾਦੀ ਮਾਂ→ ਦਾਦੀਮਾਂ

In the above two word substitutions ਜਿਸ ,ਦਾ,ਦਾਦੀ,ਮਾਂ are four different words. Sometimes these errors give rise to real word e.g. ਉਸ ਦੇ→ਉਸਦੇ, ਜਿਸ ਦੇ→ਜਿਸਦੇ. Words ਉਸਦੇ,ਜਿਸਦੇ are two valid words. The percentage of run on error is found to be 5.20% in single and 0.61% in single-error misspellings.

1.1.6 Split word errors

Split Word Error[2] : This is Opposite of Run-on error when there is some extra space is inserted between parts of a word. The error can be removed by removing the extra space. e.g. ਸਕੂਲ→ਸ ਕੂਲ, ਦੀਵਾਰ→ ਦੀ ਵਾਰetc.

Sometimes these errors give rise to more than one real word errors e.g. ਉਸਦੇ→ਉਸ ਦੇ, ਜਿਸਦੇ→ਜਿਸ ਦੇ. Words ਉਸ,ਦੇ,ਜਿਸ,ਦੇ are four valid words. The percentage of split word error is found to be 2.32% in single and 0.20% in multi-error misspellings.

5.2 Positional Analysis

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

Table 6 : Position wise distribution of misspellings

Sr. no.	Position	%age in single	%age in multiple
1	1st	13.11	12.99
2	2nd	18.98	16.25
3	3rd	26.80	23.16
4	4th	17.95	16.87
5	5th	11.30	13.20
6	6th	5.43	6.52
7	7th	3.78	4.50
8	>7th	2.65	6.50

It is generally believed that few errors tend to occur in the first letter of a word . **Pollock and Zamora[6]** found that 3.3% of the 50000 misspellings involved first letter and **Yannakoudakis and Fawthrop[7-8]** observed a first position error rate of 1.4% in 568 typing errors. The percentage of first position errors in Punjabi language is considerable. It is observed that in single error misspellings 13.11% o and 12.99% in multi error misspellings are found to be first position errors.

This rate is more than as expected. Concluded reasons are:

1. Naveen group Elements: Out of the total first position misspellings, 32.93% were the misspellings who have mistakes due to (ਸ਼,ਖ਼,ਗ਼,ਜ਼,ਫ਼,ਲ਼)i.e. where the typist has typed ਸ਼→ਸ,ਖ਼→ਖ,ਗ਼→ਗ,ਜ਼→ਜ,ਫ਼→ਫ,ਲ਼→ਲ. It means at least 32.93% of the first position misspellings are due to substitution errors. Though there are many more other substitution pairs that are also found. It is clearly signifying the probability of the substitution error at the first position. Table 7 shows the distribution of errors evolving due to each element of the group.

Table 7 : Commonly occurring confused character pairs at first Position

Sr.no.	Character Pair	% age out of total no. of first position misspellings	Cumulative %age
1	ਸ਼→ਸ	13.26	13.26
2	ਖ਼→ਖ	7.80	21.06
4	ਫ਼→ਫ	5.98	27.04
7	ੳ→ ਵ	4.43	31.47
3	ਜ਼→ਜ,	2.99	34.46
5	ਗ਼→ਗ	2.88	37.34
8	ਠ→ਲ	1.89	38.23
6	ਲ਼→ਲ	0	38.23

Note: → is showing both way substitution

2. Shifted and Unshifted modes of typing: e.g. ਠ→ਲ,ੳ→ ਵ.
3. Multiple forms for a word: e.g. ਜੇਹਾ →ਜਿਹਾ, ਵੀਚਾਰ→ਵਿਚਾਰ .The percentage of

Substitution of the above word pairs for out of the total no. of first position error misspellings is 3.15%.

5.3 Word length Effect

In English **Kukich [1990][1]** analyzed over 2000 error types ina 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. It is observed that about 56% of errors are in words of length 3,4,5(Table 8). This means words having word length of five contain maximum of errors.

Table 8 : Word Length wise distribution of misspellings

Sr.no.	Word length	%age of errors	Cumulative %age
1	1	0.1	0.1
2	2	5.15	5.25
3	3	16.75	22
4	4	20.64	42.64
5	5	21.18	63.82
6	6	16.13	79.95
7	7	8.93	88.88
8	>7th	11.12	100

Table9 : Distribution of misspellings according to word length and Type of error

Word Length Type Of Error	1	2	3	4	5	6	7	>7	Total	Cumulative %age
SE	.01	2.60	7.07	11.53	8.87	6.87	2.82	3.42	43.20	43.20
DE	.02	.40	4.60	6.56	7.77	5.43	4.04	4.79	33.61	76.81
IE	.07	2.09	3.17	2.67	2.51	2.24	0.88	1.46	15.09	91.90
TE		0.19	0.18	0.41	0.44	0.30	0.13	0.13	1.78	93.68
ROE			0.06	0.24	0.58	1.24	0.76	1.49	4.37	98.05
SWE		0.04	1.15	0.28	0.16	0.11	0.10	0.11	1.95	100.00
Total	0.10	5.32	16.23	21.70	20.33	16.19	8.73	11.40	100	
Cumulative %age	0.10	5.42	21.65	43.35	63.68	79.87	88.6	100		

5.4 Multiple Error Distribution

An analysis was also carried out for multi-error misspellings and it is observed that majority of the multi-error misspellings contain two mistakes (see Table 10).

Table 10 : Percentage of no. of mistakes in a word

Sr.no.	No. of mistakes	%age out of total no. of multi-error misspellings	Cumulative %age

1	2	81.79	81.79
2	3	11.67	93.46
3	4	5.81	99.27
4	5	0.67	99.94
5	>5	0.06	100.00

5.5 Phonetically Similar Character Error Analysis

Phonetic errors are a special class of cognitive errors in which the writer substitutes a phonetically correct but orthographically incorrect sequence of letters for the intended word. Punjabi language also contains these type of confusion characters where the typist generally

type the phonetically similar but wrong character.We have classified the phonetic errors into four categories :

1. Type 1 ਗ→ਘ,ਜ→ਝ,ਦ→ਧ,ਡ→ਢ,ਨ→ਣ, ਬ→ਭ
2. Type 2 ਸ਼→ਸ,ਖ਼→ਖ,ਗ਼→ਗ,ਜ਼→ਜ,ਫ਼→ਫ,ਲ਼→ਲ
3. Type 3 ੇ → ੈ , ੁ → ੂ , ੋ→ ੌ, ਿ → ੀ , ੰ→ ਂ
4. Type 4 ਰ→ ੍ਰ ,ਵ→ ੍ਵ ,ਹ→ ੍ਹ

It is analysed that 17.6% of the errors are due to characters belonging to Type 1.Out of the total no. of phonetic errors 59.28% are due to the Type 2 group elements.

% age of phonetically similar sounding vowel pairs is also considerable. It is concluded that about 23.83% of the misspellings contains mistakes due to Type 3 vowel pairs and 8.09% of the misspellings are due to type 4 phonetically similar pairs.

6. Conclusion

For the first time, a detailed study has been made on the pattern of Punjabi tying errors. We have done analysis based on type of errors, positional effects, first position error analysis, phonetic effects, word length effects etc. Besides the usual typing mistakes, the other reasons for majority of the misspellings in Punjabi language are due to:

1. Due to multiple forms of the same word or the non standardization of Punjabi spellings.
2. Slight difference between the pronunciation and spellings of some of the Punjabi words.
3. Due to naveen group elements.
4. Due to phonetic similarities of various consonants and vowels.
5. Borrowed words from other languages

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] Roger Mitton, Spelling checkers, spelling correctors and the misspellings of poor spellers, Information Processing and Management: an International Journal, v.23 n.5, p. 495-505, Sept. 1987.