Spelling Error Trends in Urdu

Tahira Naseem and Sarmad Hussain
Center for Research in Urdu Language Processing, FAST-NU, Lahore
tahira.naseem@nu.edu.pk, sarmad.hussain@nu.edu.pk

Abstract

Today the most accurate error correction techniques are statistical. But for low resourced languages like Urdu, where training error corpora are not available, statistical techniques are out of the question. Rule based techniques that exploit spelling error trends provide a useful alternative. The study of error patterns in a language is an essential prerequisite for designing such techniques. This paper presents two studies of spelling error trends in Urdu. The results show that alongside the already known spelling error trends common to all languages, Urdu also exhibits some language specific error patterns. The most important among them are space related errors and shape similarity based errors. They form a dominating portion of the total spelling mistakes in Urdu.

1. Literature Review

Until recently, most of the spelling correction techniques were designed on the basis of spelling errors trends (also called error patterns); therefore many studies were performed to analyze the types and the trends of spelling errors. The most notable among these are the studies performed by Damerau [1] and Peterson [4]. According to these studies Spelling errors are generally divided into two types, typographic errors and cognitive errors.

Typographic errors occur when the correct spelling of the word is known but the word is mistyped by mistake. These errors are mostly related to the keyboard and therefore do not follow any linguistic criteria.

A study by Damerau [1] shows that 80% of the typographic errors fall into one of the following four categories

1. Single letter insertion; e.g. typing acress for cress
2. Single letter deletion, e.g. typing acress for actress
3. Single letter substitution, e.g. typing acress for across
4. Transposition of two adjacent letters, e.g. typing acress for caress

The errors produced by any one of the above editing operations are also called single-errors [2]. Damerau's assertion was confirmed later by a number of researchers including Peterson [4]. The results of a study by Peterson [4] are shown in Table 1. The data sources were Webster's Dictionary and Government Printing Office (GPO) documents retyped by college students.

The rows in Table 1 correspond to four basic types of errors; the columns correspond to the two sources of data. For each data source, the number and the percentage of each type of errors is given. The last row contains total number and percentage of single errors.

Table 1. Statistics of the Four Basic Types of Errors (for English).

	GPO	Web7
Transposition	4 (2.6%)	47 (13.1%)
Insertion	29 (18.7%)	73 (20.3)
Deletion	49 (31.6%)	124 (34.4%)
Substitution	62 (40.0%)	97 (26.9%)
Total	144 (92.9%)	341 (94.7%)

Typographic errors are mainly caused due to keyboard adjacencies. The most common of these typographic errors is the substitution error (as shown in 4th row of Table 1). Substitution error occurs when a letter is replaced by some other letter whose key on the keyboard is adjacent to the originally intended letter's key. In a study referred to by Kukich [2], 58% of the errors involved adjacent typewriter keys.

According to Peterson [4] the next most common errors are two extra letters, two missing letters and transposition of two letters around a third one. The errors, produced by more than one editing operations, are called multi-errors. [2]

Cognitive errors occur when the correct spellings of the word are not known. In the case of cognitive errors, the pronunciation of misspelled word is the same or similar to the pronunciation of the intended correct word. (e.g. receive -> recieve, abyss -> abiss etc.)

In a study, referred to by Kukich [2], Dutch researchers let 10 subjects transcribe the 123 recordings of Dutch surnames, 38% of these transcriptions were incorrect despite being phonetically plausible. In another study, referred to by Kukich [2], done on spelling errors trends in students of different grades, considering only those mistakes whose frequency was greater than 5, it was found that 64.69% were phonetically correct and another 13.97% were almost phonetically correct. It was postulated that errors with lower frequency have a tendency to be less phonetic.

2. Spelling Error trends in Urdu

Two studies were performed to identify error patterns in Urdu. Due to the difference in the nature of the data and in the methodology used for studying the data, the two studies are discussed separately. Study 1 is also discussed in [3].

2.1. Study 1

2.1.1. Methodology. The data used for the study was gathered from the following resources

1. Urdu Newspapers
2. Urdu term papers typed by graduate and undergraduate university students

The data were available in the form of hard copies and were manually spell checked.

2.1.2. Results. Results of the study are shown in Table 2 (a & b). The statistics from the two sources are entered separately because the trends they exhibit are slightly different from each other. The analysis of only single-errors is given. These errors are further divided into the categories of insertion, deletion, substitution and transposition errors. For each of these categories the number of errors that were visually or phonetically similar to the actual corrections is also given. In some cases, the errors could justifiably be assigned to any of the two categories, i.e. they were both visually and phonetically similar to the intended word. In such cases one of the two factors was always seen to be clearly dominating and the error was assigned to that category. The bottom row of the table shows the total number of errors analyzed including both single-errors and multi-errors.

Table 2. Statistics of Single Edit Distance Errors in Urdu
(a) For Newspapers Text

	Newspapers Text		
	Total Errors	Visually Similar	Phonetically Similar
Substitution	75	40	12
Deletion	42	4	5
Insertion	21	2	1
Transposition	12	3	0
Total	150(91%)	49	16
	Total number of errors was 164.		

(b) For Students Term-papers Text

	Students Term-papers Text		
	Total Errors	Visually Similar	Phonetically Similar
Substitution	35	19	14
Deletion	20	4	1
Insertion	7	0	2
Transposition	5	2	1
Total	67(93%)	25	18
	Total number of errors was 72.		

2.1.3. Discussion. The results from the two sources are largely similar except that the ratio of phonetically similar errors in the term-papers text is much higher than in the newspapers text. This is because sound based errors are mainly cognitive errors, and there is little chance that a professional writer at a newspaper would make cognitive mistakes.

In the texts from both sources, the ratio of single-errors is above 90%. This matches with the results reported by Peterson [4] for English.

The data also shows that about 50% of the errors are either visually or phonetically similar to the corresponding correct words. The examples of phonetically similar errors in Urdu are پذیر /pə.zir/ → پزیر /pə.zir/ and لحاظ /li.haz/ → لحاض /li.haz/. The examples of visually similar errors are محفوظ /mɛh.fuz/ → محفوط /mɛh.fut̪/ and چبوتروں /ʧə.but̪.rõ/ → چپوتروں /ʧə.put̪.rõ/. Among these, the contribution of shape-similarity based errors is much higher. About one third of the single-errors are of this type. These errors are mostly single letter substitutions. This can also account for the greater percentage of substitution errors as compared to the percentage (about 30%) reported by Peterson [4] for English. As for English, visual similarity has never been reported to play any role in error trends.

Shape-similarity based errors cannot be cognitive in nature. There is little likelihood that a person typing the text of a language does not know the correct shapes of the letters in the language alphabet. Therefore, there should be some other explanation for this type of errors. In the authors' view the errors of this kind arise mainly for two reasons. First, the professional typists, when given a typing assignment, are provided with hand-written draft of the text that they have to copy. In this situation the typists tend to type the text as it looks without giving much attention to its meaning and as a result visually similar letters are confused for each other. Second, when a mistake of this kind is made either due to the above-mentioned reason or for some other reason like keyboard adjacencies, it goes undetected by the person typing because of its visual similarity.

Another reason for the greater number of substitution errors can be the use of shift key for typing. Many letters in Urdu are typed with the ***shift key*** pressed (due to greater number of letters in Urdu alphabets, 41 in total), so a letter might be replaced with another letter, if same key is used for typing both of them.

Phonetically similar error can be considered cognitive. These errors in Urdu are mainly caused due to ***homophone Characters***. Homophone characters are those characters, which represent the same sound. In Urdu, the number of homophone characters is relatively greater compared to English. Following are listed the homophone character sets of Urdu.

ز،ذ،ض،ظ   ص،س،ث   ط،ت   ک،ق
ا،ع   ح،ہ،ھ

It was also observed that in Urdu word initial errors are as common as are word medial or final errors. Especially word initial omission errors (مجھے /mʊ.ʤʰe/→ جھے/ʤʰe/, اسلامی /ɪs.lɑ.mi/ → سلامی /slɑ.mi/) are very common. Moreover phonetics based substitution errors (زینت /zi.nət̪/→ ذینت /zi.nət̪/, زیب /zeb/ → ذیب /zeb/) are as common word initially as they are word medially.

Another interesting observation regarding these errors was that 25% of these were real word errors i.e. they resulted in valid Urdu words. For example:

/slɑ.mi/ سلامی <– /ɪs.lɑ.mi/ اسلامی
/mɑ.hin/ماہین <– /mɑ.hɪ.rin/ ماہرین
/ t̪ʊr.ki/ترکی <– / t̪ə.rəq.qi/ ترقی
/ʧʰut̪/چھوٹ <– /ʤʰut̪/ جھوٹ

It was also found that 5% of the typing errors were space related i.e. they involved insertion, deletion, substitution or transposition of space character.

2.2. Study 2

2.2.1. Methodology. The data for this study was taken from a corpus of Urdu Text (1.7 million words) developed at CRULP. The corpus was spellchecked automatically and the errors were analyzed manually. The study contains the analysis of only non-word errors. Automatic detection of real word errors requires sophisticated algorithms which are not available for Urdu at present time.

2.2.2. Results. The results of corpus data study are presented in Table 3 and Table 4. Table 3 divides the error into space related errors and other errors. Table 4 shows the statistics of single errors with in the non space errors.

Table 3. Comparison of the Space Related Errors with Other Errors.

Non space errors	239	24.51%
Space omission	672	75.49%
Space insertion	53	
Space transposition	11	
Total Errors	975	

Table 4. Non Space Errors' Profile

Deletion	43	17.99%
Insertion	49	20.50%
Substitution	109	45.61%
Transposition	17	7.11%
Diction Variation	21	8.79%
Total (non space errors)	239	

2.2.3. Discussion. Statistics regarding four basic types of errors are again in agreement with previous studies, but the major difference is the large number of errors due to space omission and space insertion. This type of errors could not be captured through the manual study since such mistakes, most of the times, make no change in the visual form of the word, while for error analysis of corpus, the corpus was first programmatically tokenized on spaces and punctuation marks in order to separate words. Due to inappropriate use of space, too many run-on and split-up words were found, 75% of the total 975 Non-Word errors was due to missing or wrongly inserted space. Space omission is much more frequent compared to space insertion; perhaps because we always want to minimize typing effort therefore spaces are omitted intentionally but inserted either mistakenly or of necessity. This later happens mostly in the case of compound words when a space is inserted in the middle of a word just to prevent the joining of two characters within the word that are supposed to be separate. Consider the following examples:

بذله سنجی   گردن زنی

If the spaces are removed from within the words they will become

بذلہسنجی   گردنزنی

which is not correct.

Space omission error occurs because in Urdu writing there is actually no gap between words; separate words are just not joined with each other. When typing, two adjacent words do not get joined if the last character of the first word is a non-joining character, even if no space is inserted between the two words. In such situations if the space is omitted it won’t cause any noticeable difference. For example in the following sets of words there is no space between the words within a set but it seems visually quite alright.

اپناگھر   میزپرکتاب

So it can be inferred that space related errors are not actually errors, because they do not cause any observable misspelling. They are not a problem for the reader (the difference in statistics of the two studies is a proof for this implication); they are just a problem for a spellchecker or any other computational application that needs to tokenize Urdu text. And any spellchecker that does not properly tackle the errors of this kind will give too many false alarms.

3. Conclusion

From the studies presented in this paper it can be concluded that in Urdu, spelling errors exhibit a couple of script specific trends that are not found in the studies of error trends of English. One of these is the frequent occurrence of substitution errors caused due to the shape similarity of the letters in Urdu alphabet. The other is the omission of spaces at word boundaries. It can be assumed that these results will also apply to other languages that are written in Arabic script. These results imply that the existing rule based spelling correction algorithms may not be as effective for Urdu, and for Arabic script based languages in general, as they are for roman script languages. They might require modifications to cater the script specific issues of spelling errors.

4. References

[1] Damerau, F.J. A Technique for Computer Detection and Correction of Spelling Errors. *In Communications of ACM*, Vol. 7, No. 3, pp. 171-177, March, 1964.

[2] Kukich, K. Techniques for automatically correcting words in text, *ACM Computing Survey*, Vol. 14, No. 4, pp 377-439, December 1992.

[3] Naseem, T. And Hussain, S, A Novel Approach for Ranking Spelling Mistakes in Urdu, to appear in *Language Resources and Evaluation*, in June 2007.

[4] Peterson, L.J. A Note on Undetected Typing Errors. *In Communications of ACM*, Vol. 29, No. 7, pp. 633-637, July, 1986.