J. Appl. Environ. Biol. Sci., 4(7S)185-192, 2014

© 2014, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Urdu Word Processor – Standards and Guidelines

Uzair Muhammad, S. Tahir Ali Jan, M. Shahid Sultan, M. Naeem

Faculty of Computing, Riphah International University at Islamabad

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

Transliteration and word processing in Urdu is becoming under limelight in the last decade. Though there has been an immense addition of word processing features in contemporary word processors for English but as far as Urdu word processing is concerned, the researchers are seen mostly at the side of finding single aspect resolutions or optimization of Urdu language processing problems but tactlessly they give their integration (as one single text editing solution) a very less significance. Integrated solutions for the fulfillment of modern Urdu content editing needs are found to be very fewer and feeble at present. Because of having no compact word processing solution to retort present-day Urdu word processing needs, there is a great need for an Urdu word processor that could be able to justify the contemporary text editing needs of people who can read and write Urdu language.This paper presents features and requirements for a state of the art Unicode based Urdu word processor described as per the present day necessities. It also considers the comparative analysis of the present day word processing solutions, their blemishes, and their strengths over each other and what more can be done in this area to promote Urdu word processing would be described in this paper as well.

KEYWORDS: Urdu word processing, Natural Language Processing, Guidelines for Urdu word editor, Urdu Editor.

1 INTRODUCTION

According to historical analysis, Word Processing first gained popularity in 1971 and people compelled to think about the paperless office in order to make their repetitive office work easier and faster [17]. People who started this concept had a primary focus on English language processing features and as the time passed more and more solutions got added into existing ones to improve them leaving other languages blurred.

At present, Urdu is spoken by 63.4 million people all over the world [23]. In Pakistan, Urdu is an official language and only 20% of the total literate population of Pakistan understands English [15] therefore, the word processor must be in Urdu language so as to get fully comprehended by the native people and they could be able to take maximum gain of the current technology.

Urdu being a subset of Arabic language has been of great interest for researchers and scholars since its development. Many problems were answered and still being answered by many researchers encasing various facets of it Today in this technological era, where office automation [5][6] has caused a stir and unabridged office work has come in just a single contraption, for the people who have a command over Urdu language only prefers text editors to be in their native language as this can enhance their productivity for their routine work in their respective disciplines. Many efforts were made in this regard and companies came with their proprietary solutions. These solutions are still being used by people but certain limitations offered by them deter accomplishing their required goals at different levels of usage. There emerges a great demand for modern word processor that could accommodate modern word processing requirements of the people whose primary language is Urdu.

Intrinsically, this research is being prompted to standardize the features of a present-day word processor for Urdu language. With some of the common features found in every existing word processor,

* **Corresponding Author:** Uzair Muhammad, Faculty of Computing, Riphah International University at Islamabad. Email: uzair.muahmmad@riphah.edu.pk,

we are also mentioning some of the contemporary features which are considered to be assimilated in every modern word processor.

Next section gives an idea of the domain under the title background and related work and give and idea of the state of the art in literature and industry. Section 3 proposes some features and standards of Urdu Word Processor (UWP). Discussion section and comparison of the features goes as section 4. Section 5 mention the future direction and last section is the conclusion.

2 BACKGROUND AND RELATED WORK

Urdu being a subset of Arabic language has been a focus of research since researchers got into the need of merging and digitizing Urdu language support with computing. With the influx of Unicode (2-byte character encoding standard), character mapping for Urdu language got evolved and this revolution brought the courtesy of researchers to reform or optimize the existing solutions and finding solutions to the existing problems. Few classic problems and their existing solutions depicting various aspects of text editing paradigm are as follows: OCR (Optical Character Recognition) for Urdu Language has got a tremendous effort in the last few decades and still is getting on peak. One research proposed a solution irrespective of the Fonts or special scripts for both offline and online platforms by utilizing their own algorithmic approach instead of the classic approach to this which is Artificial Neural Network [30]. Another research developed a solution for the conversion of Urdu Nastaliq Font into Roman Urdu using OCR. In that proposed technique they are segmenting each character and matching each to their corresponding character already saved in database [14]. On the other hand, many researches have been proposed regarding English to Urdu translation amongst which one is an Expert System Based approach. This solution provides an algorithmic approach which promotes less usage of dictionary lookups [29].

Still these out of many solutions are being optimized and researchers are trying to improve the digitization of Urdu language in order to make it easy for the future work. Here the problem is that the researchers are taking each of the solutions in isolation. In other words, no effort has still been initiated to integrate the basic solutions in order to make a compact and contemporary text editing solution. We are trying to promote this effort so as to formulate basic text editing needs and integrate them into one single solution.

Quite a few text editors are available for Urdu-like Unicode compatible languages in the literature. One editor has been mentioned in [22] for Urdu and Burushaski, and the other one in [29] with the name “AGHAZ” with algorithms for translation of English into Urdu. The Former editor has dwelled its roots in literature with the intent to support and preserve the Burushaski language spoken in Northern areas of Pakistan. This editor provides only few of the basic features of text editing making it really necessary to improve it. The latter research provides a single aspect of text editor which is “Translation from English to Urdu”. This research instead of providing a compact solution to content editing needs, lays foundations for single aspect investigation.

There are a few industry products available. The existing solutions would be scrutinized on the basis of the pros and cons they offer individually.

2.1 InPage® Professional

Inpage V3.0 (stable release) released in 2008, by Concept Software Pvt. Ltd. Its ever first version was released in 1994. The total number of registered users is more than 1,000,000 in Pakistan [10]. Among the other features it provides Spell Checking, conversion into PDF, Unicode support and Object operations. Since this is proprietary software it has a price of US$ 394 (Concept Software, 2013) which makes it costly for most home users. Similarly it still lacks major features which are required for today’s user.This is the sub heading of the paper. This is the sub heading of the paper. This is the sub heading of the paper. This is the sub heading of the paper.

2.2 UrduEditor

UrduEditor V8.5, released by SummitSoft. The total downloads of this editor are 116,714 (till 11/03/2014) [27]. It has Unicode support, virtual Urdu On-screen keyboard, with user interface in Urdu. However it costs US$ 22.

2.3 Nigar Unicode

Nigar Unicode V2.0.1.5, released by AJSoft. The total downloads of this editor are 30,850 (till 11/03/2014) [3]. It is a Freeware and provides Unicode sustenance, spell check functionality and exporting as image option as well. The odd side of it is that it has no option to export file as PDF (Portable Document Format), no object operations supported by it, and no blogging at social media sites.

All these editors, available both in industry and literature, provide some basic and advance level functionality, however, still they are lacking too many features that are requirement of any text editor of the day to cover the needs of today's user. After a basic comparison of these tools or editors, we can comprehend that no single solution afford all compulsory features in it and hence, this may also limit one's ability by lacking basic features. So we need single integrated text editing solution that could live up to both industry as well as literature echelons.

3 Proposed Standards and Features for Urdu Word Processor

After having a closer look and thorough investigation of different Urdu editors and tools we compiled and present a list of requirements and features for Urdu Word Processor. Authors are currently developing an Urdu Word Processors that will provide all the features that have been identified in this study. This section describes their requirements for the features of the proposed solution Urdu Word Processor (UWP).

Following Table will look around the basic features that are considered to be present in almost all of the text editing softwares and our proposed solution would integrate them all as well along with other contemporary features being discussed one by one ahead.

Table 1. Common Features in most Word Processors

Group	Features
File	New, Open, Save, Save As, WYSIWYG (Print, Print Preview, Print Settings), Exit
Edit	Undo/Redo, Copy, Paste, Cut, Select All, Find, Find Next, Replace, Replace All
View	Zoom-in/Zoom-out, Status Bar, [any other bars/tools]
Insert	Page break, Time and Date, Symbol, Table
Format	Align[Left, Right, Center, Justify], Text Decoration [Underline, Bold, Italic], Bullets, Increase/Decrease Indentation
Text	Word Statistics, Spell Checker
Customization	GUI Customization
Font	Size and Face

With these common features, following are requirements/features that are presented as standard for any UWP. Most of these features have been incorporated successfully into the UWP, which is under development. This project, with all the features and standards mentioned in this paper, will be completed within a couple of months.

3.1 Unicode compatibility

An Urdu Word Processor (UWP) should be able to handle and process Urdu Unicode characters [32]. They are mapped from Arabic range (0600-06FF) Hex, extended Presentation 'A' ranges (FB50-FDFF) Hex and Presentation 'B' ranges (FE70-FEFF) Hex along with Persian, Sindhi, Pashto, and Kurdish etc. languages of Iran, Pakistan, and India. The [18] *"Urdu Computing Standards: Development of Urdu Zabta Takhti"* would be used for Urdu character sets and keyboard design layout.

For a contemporary word processor, it has to be compatible with Unicode, which covers more than 100 scripts [2]. Modern web and computer languages support Unicode and more than 382 Urdu fonts online [28] that support Unicode; so the Urdu word processor must support Unicode.

3.2 Compatibility with Rich Text Format (RTF)

We intend to use Rich Text Format (RTF) (which supports Unicode, needs to be escaped [7] primarily which can be converted to various text formats and markup languages like HTML, WordML, and XML etc.

3.3 Email Support

We reckon that there is an immense need that the modern word processors must be able to send the document via email right from within the editor. It is therefore becoming a core requirement to integrate the email functionality inside the GUI of the Urdu Word Processor as the patent [11] by Microsoft suggests to make it uniform with other editors.

3.4 Bi-lingual User Interface (UI)

English is not the first language in Pakistan and therefore typical speakers of Urdu find it difficult to comprehend menu commands and other message boxes that are in English.It therefore become one of the core requirement to introduce a bi-lingual User Interface to facilitate users. For this purpose it is recommended to follow Microsoft's Urdu Style Guide [12] and therefore we would be following it in our underdeveloped UWP.

3.5 Nastaliq Font as Urdu UI

Urdu language has many writing scripts but commonly used script in Urdu language is Nastaliq [16] therefore, we would employ Nastaliq font style for the Urdu User Interface so that user finds a familiar font and a familiar User Interface which will in return increase the productivity of the user.

3.6 PDF file Generation

PDF (Portable Document Format) was first introduced by Adobe in 1993 [1], PDF has become an open standard for electronic document exchange maintained by the International Organization for Standardization (ISO). We intend to use "iTextSharp" library to generate the document to PDF.

3.7 Equation Creator

Mathematical Equations are important in different reports and research papers. Even most of the mathematics books at elementary level are in Urdu. Therefore support to write Mathematical Equation in UWP is a requirement. We intend to integrate "Math Editor Mini 1.0" [19] which would generate the equation as an image to be inserted into the editor.

3.8 Publish to Social Media

With over 10-milion users [24] users from Pakistan, there is a need for automated Urdu micro-blogger for Social Media Websites. UWP shall be able to facilitate its user to publish content to social media website e.g. Facebook.

3.9 Publish to WordPress

Besides micro-blogging there is a considerable amount (75,187,962) of WordPress powered sites around the World Wide Web [31], to fulfill the needs of Urdu bloggers we would be integrating the XML-RPC API for posting articles directly from within the Urdu Word Processor.

3.10 Mail-merge

MicroPro started shipping WordStar in June 1979. After completing most of the code, Barnaby was joined by Jim Fox, who helped complete the code and wrote the installation program. Barnaby and Fox continued to improve WordStar through several interim versions. A major innovation was the development of the Mail Merge program, which allowed the insertion of names and addresses into a WordStar document from a separate file [5][6].

Mailing same content to a number of receivers can be a cumbersome job which needs to be automated. We will propose a GUI inside the anticipated editor through which user can either write letters for different receivers with or without envelope labels or user can send an email using this option.

3.11 Text Art

Using typography and text art we better express our feelings and motives, therefore, for better and beautiful headings we will offer this feature by inserting the heading as an image for the time being because of the lack of technology within our resources.

3.12 Word Statistics

Statistics is very important for any word processor, the total number of words, number of lines, and number of pages etc. is to be shown to the user. This feature is not something ordinary that text editors takes into account rather this is integrated in order to provide aid to the user as he could check the number of words, lines, or characters he is supposed to write to fulfill the restriction imposed on him/her by some authority to limit the words to some boundary line. Proposed "Urdu Word Processor" is expected to accomplish this task by showing total number of characters and number of lines in the "Task-bar" area and a complete information in a dialog-box.

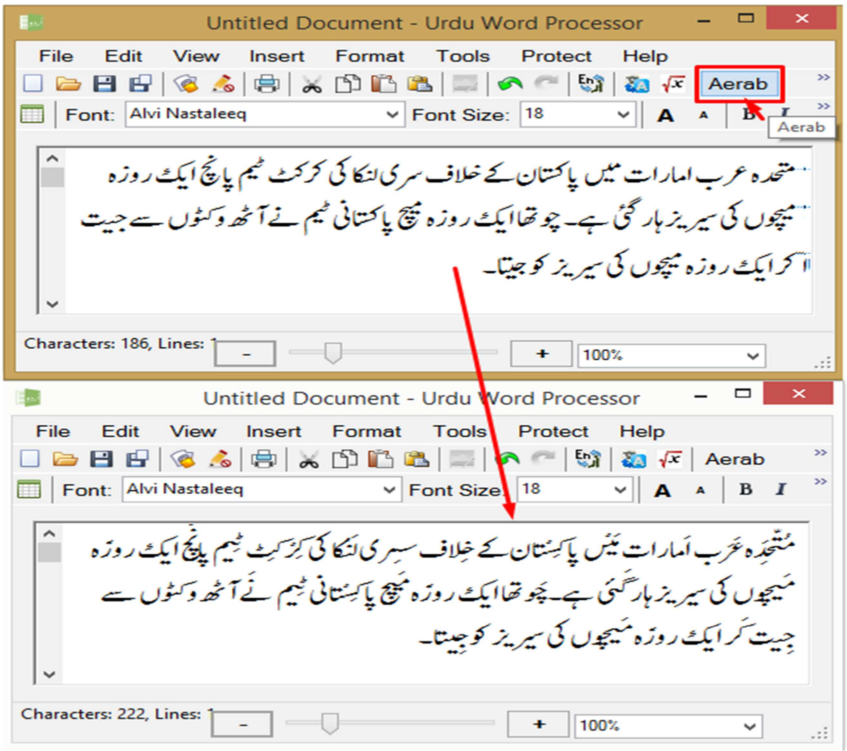
3.13 Appearance of a certain word in document

It is sometimes crucial to limit certain words in speech or in article; we sometimes cannot directly count the number of times a certain word is being used in such article/essays. To solve this problem we will offer a method to count a word appearing inside the document by showing the figure/number with highlighting all the occurrences of the word.

3.14 Automatic Aerabification (Diacritization) of Urdu Text

Automatic Diacritization[1]is the unique feature would be provided by Urdu Word Processor as shown in figure 1. We first intend to gather more than 200,000 words from different Urdu blogs and news websites and then we will be using "Urdu to ASCII Transliteration system" [8] (which offers Diacritization of Urdu words) to store the diacritized words in the SQLite3 database.

We will use database because the "Urdu to ASCII transliteration system" was taking too long to process words which apparently would make the process difficult for the end users.



Font: Alvi Nastaleeq Font Size: 18

متحدہ عرب امارات میں پاکستان کے خلاف سری لنکا کی کرکٹ ٹیم پانچ ایک روزہ
میچوں کی سیریز ہار گئی ہے۔ چوتھا ایک روزہ میچ پاکستانی ٹیم نے آٹھ وکٹوں سے جیت
کر ایک روزہ میچوں کی سیریز کو جیتا۔

Characters: 186, Lines: 1

Font: Alvi Nastaleeq Font Size: 18

مُتّحِدَہ عَرب اَمارات مَیں پاکِستان کے خِلاف سِری لنکا کی کِرکِٹ ٹیم پانْچ ایک روزہ
مَیچوں کی سیریز ہار گَئی ہے۔ چوتھا ایک روزہ مَیچ پاکِستانی ٹیم نے آٹھ وکٹوں سے
جِیت کَر ایک روزہ مَیچوں کی سیریز کو جِیتا۔

Characters: 222, Lines: 1

Figure. 1. Automatic Aerabification.

3.15 Urdu to Roman-Urdu Conversion

Urdu without diacritics is hard to convert to Roman-Urdu, therefore, we will use the same database to map Urdu words to Diacritized Urdu words and then to Roman. The Urdu words would be converted to

[1] Diacritization is the addition of normally unwritten letters which if written with words describe their pronunciations like Zabarَ, Zairِ, Paishُ, Madd~ etc.

Roman-Urdu with the help of table mentioned in "Conversion of Urdu nastaliq to Roman-Urdu using OCR,"[21].

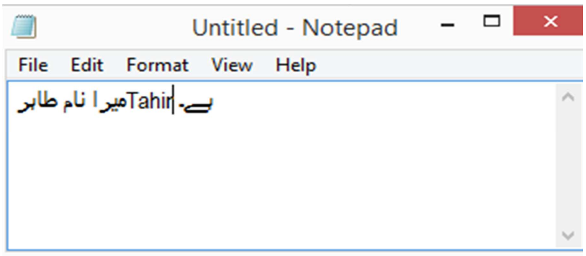
3.16 Print and Print Preview of the document

The ability to Print and Preview a document is an essence of a Word Processor which we reckon that it should be implemented.

3.17 Phonetic Urdu Keyboard input

Phonetic Keyboard Layout [9] [26] is very popular and easy to use [4] we decided to program the input facility into the Extended Rich Text Box so that regardless of keyboard layout and language being installed, the user should be able to type and process Urdu. For this we will extend a regular RichTextBox, so that when user types “s” and the input language set to Urdu, “س”will be typed, and when user types “S”, “ص”will be inserted.

Here, it is worth mentioning that the RichTextBox or any Text Box, the text direction should be set to “Right-to-Left” or else the text line will be dis-positioned when English word is being typed inside Urdu, as figure 2 depicts:



میرا نام طاہرTahir ہے.

Figure. 2. Urdu Text Dispositioned.

3.18 On-screen Virtual keyboard

A QWERTY virtual keyboard was designed for English while for the convenience of Urdu Users we would use the famous Phonetic Keyboard Layout [9], our virtual keyboard would also contain a shortcut button for inserting Symbols and a button for Urdu Keyboard Layout customization.

3.19 Urdu keyboard customization

There come certain occasions when user wants to change a specific key up to his/her ease and desire. We want to provide this ease of change to increase the user productivity.

3.20 Export document as a Microsoft Word Document

To facilitate our users so they can further use the document in Microsoft Word we used an API to export the Urdu document in Word document .doc format.

3.21 Export document as HTML document

Urdu Word Processor is not just for publishing and creating text, it can also be used as an HTML editor. We will provide facility to save the document as formatted HTML file and also to read HTML files.

3.22 Export document as a PNG image

Already in InPage, many of the times, the Urdu text is needed as an image to be published and to be used in specialized image processing software for flyers etc. We intend to provide this feature so that user could be able to export the formatted text as a PNG (portable network graphic) image.

3.23 Bi-directional compatibility with the older versions of InPage

InPage is a famous Urdu Text Editor with 1-milion users in Pakistan (InPage). Previous versions of InPage are also being used very common in Pakistan. Sadly the previous version has only one flaw, it has its own legacy encoding scheme.

We intend to offer three features to interact with InPage:

1. InPage text can be imported using Paste Special in Urdu Word Processor.
2. InPage legacy file (.inp) can directly be opened in Urdu Word Processor.

Unicode code text from Urdu Word Processor can be converted to the legacy code of InPage and pasted into InPage.

4 DISCUSSION AND ANALYSIS

The features mentioned in table 1 are common in almost all word processors and urdu word editors/processors are not an exception. Majority of the features proposed in section 3 are not included in any of the editors mentioned above. The frequency of releasing urdu editors/processors are also very low as stated in section 2. All the editors take couple of years to release a new version. For example InPage ®, with more than 1,000,000 registered users in Pakistan [10], is a major stackholder in the domain. Its lattest version is 2013, which is released after 3-4 years, still lacking majority of the features mentioned above. The features mentioned in table 1 (the basic one) and the one mentioned in section 3 (the state-of-the-art) would jointly make an editor to be called state of the art and would fulfill current day needs of office and home users of Urdu language.

5 Future work

After congregating and assembling the cluster of requirements mentioned above, we intend them to be more refined and extended with the passage of time. Our motive is to put those aforementioned features into an Urdu Word Processor to live up to the users' present-day needs regarding word processing. Ahead in near future, we would be developing word processing application that would cater all these features hence, would evolve a new breed of word processing application in computing.

6 Conclusion

One cannot deny the need of word processors for the native language as a means to promote culture and literature. The features we have described for the upcoming word processors are highly demanding and required by contemporary users. The mentioned features are supposed to be the core elements of any modern word processor and is a must have requirement for the Urdu Word Processor as well.

REFERENCES

1. Acrobat/PDF History. (2012, December 27). (Adobe Systems Inc.) Retrieved February 3, 2014, from http://acroeng.adobe.com/wp/?page_id=20.
2. Scripts and Languages. (2013, 9 14). (Unicode Consortium) Retrieved January 31, 2014, from http://www.unicode.org/cldr/charts/latest/supplemental/scripts_and_languages.html.
3. AJ Soft. (2014, March 11). Nigar Unicode. Retrieved March 11, 2014, from CNET DOWNLOADS: http://download.cnet.com/Urdu-Nigar-Unicode/3000-2079_4-10871699.html.
4. Becker, D., &Riaz, K. (2002). A study in Urdu corpus construction. Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12, Association for Computational Linguistics, 12, 1-5.
5. Bergin, T. J. (2006). The Origins of Word Processing Software for Personal Computers: 1976-1985. IEEE Annals of the History of Computing, 32-47.
6. Bergin, T. J. (2006, October–December). The Origins of Word Processing Software for Personal Computers: 1976-1985. IEEE Annals of the History of Computing, p. 16.
7. Burke, S. M. (July 2003). RTF Pocket Guide. O'Reilly Media.
8. Center for Language Engineering. (2011, August 23). Urdu to ASCII Transliteration System. (Center for Language Engineering) Retrieved February 3, 2014, from http://www.cle.org.pk/Downloads/langproc/Transliteration%20Tools/Transliteration%20tool%20for%20Windows.zip.
9. Center of Language Engineering. (2010, September 01). CRULP Urdu Phonetic Keyboard Layout v1.1 for Windows. (Center of Language Engineering) Retrieved February 3, 2014, from http://www.cle.org.pk/software/localization/keyboards/CRULPphonetickbv1.1.html.
10. Concept Software. (2013, 12 28). Urdu Inpage. Retrieved 03 11, 2014, from Inpage.com: http://www.inpage.com.

11.Darren A. Apfel, R., David M. Buchthal, D., Steve Rayson, S., Andrew G. Carlson, R., Christopher Antos, K., & Hai Liu, R. a. (2002, June 11). Patent No. US 6,405,225 B1. United States.

12.Durrani, D. A. (2011, February). Style Guides. Retrieved February 3, 2014, from http://www.microsoft.com/language/en-us/styleguides.aspx

13.Durrani, N., & Hussain, S. (2010). Urdu Word Segmentation. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 528-536.

14.Faiza Iqbal, A. L. (n.d.). Conversion of Urdu Nastaliq to Toman Urdu Using OCR. 4.

15.GUL, S. (n.d.). Dilemmas of Localization in Asia- A Case Study on Localization in Pakistan. Center for Research in Urdu Language Processing (CRULP).

16.H. Sarmad, D. N. (2005). Survey of Language Computing in Asia. Center for Research in Urdu Language Processing National University of Computing and Emerging Sciences, no. 19, 132-140.

17.Haigh, T. (2006). Remembering the Office of the Future: The Origins of Word Processing and Office Automation. IEEE Annals of the History of Computing(4), 6-31.

18.Hussain, S., & Afzal, M. (2001). Urdu computing standards: Urdu zabtatakhti (uzt) 1.01. Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, 223-228.

19.Imran, K. (2013, April 21). OOP in the Real World - Creating an Equation Editor. Retrieved February 05, 2014, from http://www.codeproject.com/Articles/522345/OOP-in-the-Real-World-Creating-an-Equation-Editor

20.InPage. (n.d.). (Concept Software Private Limited) Retrieved February 5, 2014, from http://inpage.com/

21.Iqbal, F. L., Kanwal, N., & and Altaf, T. (2011). Conversion of urdunastaliq to roman urdu using OCR. In Interaction Sciences (ICIS), 2011 4th International Conference(4th), 19-22.

22.IrfanQadirBaig, M. S. (n.d.). Multi Language Text Editor for Burushaski and Urdu through Unicode. World Academy of Science, Engineering and Technology (p. 4). International Journal of Computer, Information Science and Engineering Vol:1 No:3, 2007 .

23.Lewis, M., P., Gary F., S., & Charles D., F. (2013). Statistical Summaries. (Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas: SIL International.) Retrieved January 31, 2014, from http://www.ethnologue.com/statistics/size

24.Nasir, S. (2013, September 25). Pakistan crosses 10 million Facebook users. Retrieved February 4, 2014, from http://tribune.com.pk/story/609177/pakistan-crosses-10-million-facebook-users/

25.Rashi, R., &Latif, S. (2012). A Dictionary Based Urdu Word Segmentation Using Maximum Matching Algorithm for Space Omission Problem. Asian Language Processing (IALP), 2012 International Conference, IEEE, 101-104.

26.Rehman, B., & Qureshi, T. (2011). Urdu as Interface Design Language - A Novel Approach. Electronics, Communications and Photonics Conference (SIECPC), 2011 Saudi International. IEEE, 1-6.

27.SummitSoft. (2014, 03 11). UrduEditor. Retrieved 03 11, 2014, from CNET DOWNLOADS: http://download.cnet.com/Urdu-Editor/3000-2352_4-10062075.html

28.urduweb.org. (2013). Jameel Font Showcase. (UrduLog) Retrieved January 31, 2014, from http://font.urduweb.org/

29.Uzair Muhammad, K. B. (2005). AGHAZ: An Expert System Based approach for the Translation of English to Urdu. World Academy of Science, Engineering and Technology 12 2005, (p. 5).

30.Wahab, S. S. (n.d.). Optical Character Recognition System for Urdu. 5.

31.WordPress.com. (2014). WordPress Sites in the World. (WordPress.com) Retrieved February 4, 2014, from http://en.wordpress.com/stats/

32.Zia, K. (1999). Towards Unicode Standard for Urdu. the Proceedings of 4th Symposium on Multilingual Information Processing (MLIT-4).