Assessing Urdu Language Support on the Multilingual Web

Huda Sarfraz **Aniqa Dilawari** **Sarmad Hussain**
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology
firstname.lastname@kics.edu.pk

Abstract

This paper presents an assessment of the support available for developing web content and services in Urdu language. The paper presents the methodology designed to conduct the assessment and presents the results in its context. The paper specifically analyzes Urdu language support from aspects of character set and encoding, font support, input methods, locale, web terminology, HTML, web technologies and advanced application support. The methodology proposed can also be extended and used to evaluate support of other languages for online publishing.

1 Introduction

The web is playing a pivotal role in bringing information to the populations around the world. Though a significant amount of the content on the web is in a few languages (Internet World Stats, 2010), the web has started becoming increasingly multilingual. With the linguistic and cultural diversity come specific requirements. For example, the HTML tags used in formatting online text are largely centric to Latin script and formatting, and would need to be revised to cater to other languages using other scripts. This is evident from the fact that underlining tag <ul> causes text in languages using Indic scripts, which have a top-line instead of a base-line, to become unreadable (Lata, 2010) and is therefore not applicable.

As more languages come online, it is important to comprehensively assess the support provided for them on the multilingual web.

Urdu is one such language, with over 100 million speakers in Pakistan, India and other regions (Lewis, 2009). It is the national language of Pakistan and state language of India. Urdu uses the Nastalique writing system, which is highly cursive and context dependent, and is therefore very complex (Hussain, 2003; Wali et al., 2006). Though basic support for publishing Urdu online exists, a comprehensive analysis of the requirements and existing support is needed so that gaps can be identified and addressed.

The need for local language computing was recognized and incorporated into the IT Policy of Pakistan in 2000. The Ministry of IT has been funding research and development in this area since then. Due to these and other similar efforts, Urdu in Pakistan falls within the category of moderately localized languages, with fairly active academic and research programs, fairly mature standards, basic applications and reasonable work in advanced applications (Hussain et al., 2008a).

This work defines a methodology to analyze and assess the support for Urdu language on the multilingual web, and presents the results obtained using the methodology. The work has been undertaken to develop a consistent framework for online publishing, especially for citizen services being provided by the government in Urdu in Pakistan.

Section 2 gives an overview of related work, followed by Section 3 which gives the assessment methodology and the results obtained for Urdu. The paper then concludes with some recommendations in Section 4.

2 Related Work

With an increasing focus across the globe on the creation of indigenous and local language content, efforts are also being made to enable support for multiple languages.

The World Wide Web Consortium (W3C) states one of its primary goals is to make the benefits of the web “available to all people, whatever their hardware, software, network infrastructure, native language, culture, geographical location, or physical or mental ability” (World Wide Web Consortium, 2011). The W3C Internationalization Activity in particular collaborates with W3C working groups and other organizations to enable web technologies for use with different languages, scripts and cultures.

Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

As the online content becomes increasingly multilingual, there are multiple initiatives which are looking at existing and emerging challenges. The recent Multilingual Web project of W3C is one of the initiatives in this regard, organizing four public workshops for participants to learn about existing standards, to assess the current situation and to identify the gaps to be addressed.

In the issues being identified, Froumentin (2010) highlights web usage, and notes that 50% of the world population has access to the web but does not use it. One of the reasons cited is that native languages are not supported. Lata (2010) assesses web technology in the context of Indian languages. India has rich language diversity and Lata (2010) reports 122 major languages and 2371 dialects according to the census in 2001. The report presents a survey of multilingual web based services in India, looking at complexities in various scripts. Standardization issues are separated into three categories, with input, encoding and display issues making up the core category. The middleware category includes HTML, CSS, web accessibility and mobile web issues.

Constable and Nelson (2010) also notes character sets as posing problems in client-server interaction. It underlines the importance of "global-ready" HTML/CSS, and also formatting preferences specific to certain cultures.

Apart from the Multilingual Web Project, there are also other initiatives which have been focusing on language support on the web. The PAN Localization project is one such example, which focuses on assessing and developing the technology for enabling multilingual computing. Through the project, Hussain et al. (2005) report an analysis of the status of language computing support for 20 Asian languages. There has also been more detailed work on specific languages (e.g. PAN Cambodia 2007).

The W3C Internationalization Tag Set is an endeavor in the same direction. It is a W3C recommendation to help internationalize XML based content (Lieske and Sasaki, 2010).

3 Urdu Language Support Assessment

Assessing a language for online publishing would require to investigate its support at multiple levels. These include support in international standards, national recommendations based on international standards and frameworks, and availability of tools and applications for basic localization. Finally, for effective use, intermediate and advance application support is also desired.

This section presents the analysis of Urdu language support on the web from nine different aspects including (i) character set and encoding, (ii) input methods, (iii) fonts, (iv) collation sequence, (v) locale, (vi) interface terminology, (vii) formatting based on HTML tag set, (viii) support through web technologies, and (ix) advanced application support. Each subsection briefly describes the methodology used to analyze support for each aspect and then presents the results for Urdu language.

3.1 Character Set and Encoding

Character set and encoding support for any language is the most basic level of support needed if the script of that language is to be represented on a computational platform. The character set for any language includes basic characters, digits, punctuation marks, currency symbols, special symbols (e.g. honorifics), diacritical marks and any other symbols used in conventional publishing.

Recently Unicode (Unicode 2010) has emerged as the most widely used international standard through which character sets for different languages are enabled on the web, and in the words of Tim Berners-Lee, is the path "to making text on the web truly global" (Ishida, 2010).

When a national character set and/or encoding mechanisms exist, the next step is to ensure that the character set is appropriately mapped on to the relevant section of Unicode standard. This is easier for scripts which are based on single or few language(s) (e.g. Khmer, Lao, Thai), but becomes increasingly essential and difficult for scripts which are used for writing many different languages (e.g. Arabic and Cyrillic scripts) because in such cases the relevant subset of characters within the same script have to be identified. This becomes even more difficult for some scripts as Unicode has redundancy and ambiguity due to backward compatibility and other reasons. As more than one Unicode can be used to represent a character, termed as variants, a mapping scheme also needs to be defined to reduce the redundancies. Unicode does not stipulate all such possibilities, and a national policy has to be set to interpret Unicode in the context.

In addition to mapping, normalization is needed when a character can be ambiguously represented either using a sequence of Unicode code points or a single code point. The Unicode standard defines normalization forms in order to address this issue. However, the normalization set may not address all specific needs for a language and additional normalization tables need to be defied.

Other additional considerations include linguistic specification at national level and support at Unicode level for bidirectional behavior of a language (e.g. in Arabic script letters are written from right-to-left, but digits are written from left-to-right), character contraction (e.g. ‘ch’ is considered a single character in Slovak), case folding (Latin ‘a’ may be treated similar to ‘A’ in some contexts, e.g. to resolve domain names) and conjuncts (e.g. in Indic scripts, a consonant cluster may combine to give an alternate shape) should also be investigated for relevant languages and formalized. The Unicode standard recommends mechanisms to handle this, but it has to be determined whether sufficient support exists for this within Unicode and across various keyboards, fonts and applications.

Character Set and Encoding Assessment Methodology: The complete national Urdu character set for Pakistan, referred to as the *Urdu Zabta Takhti*, UZT 1.01, (Hussain et al., 2001) has already been defined and missing characters added to the Unicode standard (Hussain et al, 2002). However, there still no nationally accepted recommendation on which subset of Unicode from the Arabic script block (U+06XX) should be used for Urdu. Due to ambiguity in the Unicode this results in variation across different content developers on the selection of underlying Unicode characters. So a subset was defined in reference to the national character set. All normalization and mapping schemes to use this subset are also identified in the current work. The work also tested bidirectional features (Davis, 2009) for Urdu as supported by various applications.

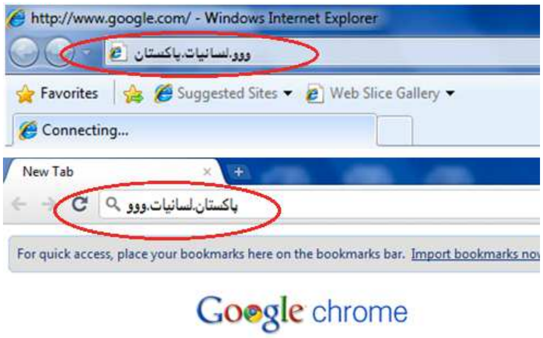
Character Set and Encoding Assessment for Urdu: The recommended character sub-set for Urdu has been defined as part of the study and is available in Appendix A. The table shows only the characters required for Urdu within the Arabic code page. Some characters have been marked as variants (V). These are characters which are not part of the Urdu character set, but bear enough similarity with particular Urdu characters that they may be used interchangeably. These have been noted because they should be mapped to corresponding characters in core set for Urdu, if inadvertently used.

Normalization and mapping schemes needed to use with the encoding are also developed and are summarized in Table 1.

Table 1. Normalization composed and decomposed forms for Urdu.

Combining Mark	**Composed Form**	**Decomposed Form**	**Unicode Normalized Form**
U+0653	U+0622	U+0627 U+0653	Defined
U+0654	U+0623	U+0627 U+0654	Defined
	U+0624	U+0648 U+0654	Defined
		U+0649 U+0654	Not Defined
		U+06CC U+0654	Not Defined
	U+06C0	U+06D5 U+0654	Defined
		U+0647 U+0654	Not Defined
	U+06C2	U+06C1 U+0654	Defined
		U+0647 U+0654	Not Defined
	U+06D3	U+06D2 U+0654	Defined

The bidirectionality analysis showed that there are some shortcomings in terms of application support. Though most of the document processing applications support Urdu properly, the newer contexts are still behind in implementation. For example, the address bar for Google Chrome (version 6.0.472.63) does not support bidirectional text, but is supported by Internet Explorer and Firefox. This support is needed for properly displaying the recently announced Internationalized Domain Names (IDNs; Klensin 2010). This is illustrated in Figure 1 below, which shows the same string ووو۔لسانیات۔پاکستان in two different browsers.



For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks no

Figure 1: Bidirectional text rendering inconsistencies in browsers for IDNs

3.2 Input Methods

Standardized input methods must be available in order to create web content and to enable users to interact with online systems and to create their own content, e.g. keyboards, keypads, on-screen keyboards, etc. All characters for the language must be supported by the keyboard layout. Any additional characters used, e.g. Latin '.' and '@' could also be supported to allow easier online access.

For many languages, a phonetic keyboard layout is possible, which allows for easier typing based on the sounds of the letters of QWERTY keyboard. Though these keyboards are normally ad hoc, they can allow for easier transition of users familiar with English keyboard to local languages and should be considered. Non-phonetic keyboards are usually based on the frequency of characters and have better efficiency, especially in the case of users who are accustomed to using them.

In deciding the keyboards to adopt, existing user base must be considered. Users who are used to an existing layout are usually reluctant to switching to a new layout even if it is more intuitive or efficient. Further, if additional characters are added to a keyboard, it is preferable for it to remain backward compatible, for adoption by existing users.

A key can represent different characters if used in conjunction with the Shift, Alt and Control keys. Though this increases the number of possibilities, increased combinations, or number of keyboard layers, significantly impact the usability of a keyboard.

Further, many languages may require additional rules, which take more than a single keystroke to generate context sensitive codes. The input method (along with the keyboard) should support such rules.

Input Methods Assessment Methodology: The current work surveyed historical typewriter layouts, starting from the first layout standardized by the Pakistan Government in 1948. Six popularly used current keyboards are also analyzed as per the framework and two recommendations are made for use based on current use.

Input Methods Support Assessment for Urdu: The CRULP 2-layer phonetic keyboard is recommended for users who are already familiar with English keyboard layouts. For new user, the Inpage Monotype layout, widely in use across the publishing industry, is recommended. However, these and other keyboards are missing '.' and '@' symbols used for web browsing and email, and thus they need to be updated. These keyboard layouts are shown in Figures 2a and b.

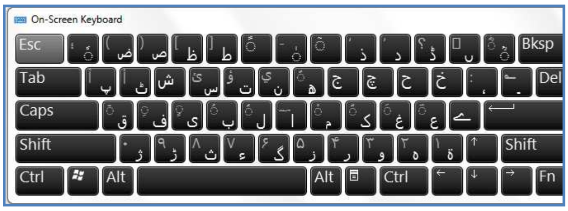


Figure 2a: Inpage Monotype layout

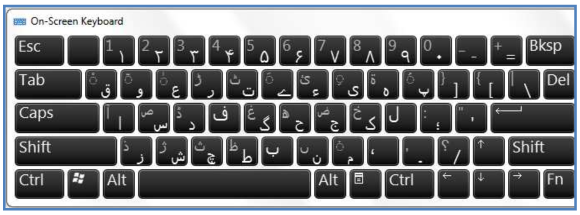


Figure 2b: CRULP Phonetic (2 Layered) layout

3.3 Web Fonts

A character is a logical entity that is visually represented through different fonts. The fonts must be based on Unicode encoding and in internationally acceptable formats, e.g. TrueType, OpenType, etc. for wider use online.

Web Fonts Assessment Methodology: In order to assess the support for Urdu, a detailed analysis of existing fonts was conducted. The fonts were analyzed in terms of the following aspects.

The character set for Urdu was subcategorized into further levels: core, secondary, tertiary and variant. The core set is minimally needed to write Urdu. The secondary set includes characters which are used in Urdu but are not part of the core set. Without these, the text is still readable, but with some difficulty. Tertiary characters were those that are used in Urdu, but their lack of support will only cause minor inconvenience. Variant characters are those that are not part of the Urdu set, but bear resemblance to core characters. If they are inadvertently used within the language, they must be supported in order to keep the text readable. This categorization was primarily done on a linguistic basis, however Google search hit counts for different character codes were used as a secondary quantitative source of evidence for this categorization. A support score was then calculated for each font being analyzed, using the scheme depicted in Table 2.

	Full Support Score	**Partial Support Score**	**No Support Score**
Primary Character	3	1.5	0
Secondary Character	2	1	0
Tertiary Character	1	0.5	0

Table 2: Scoring Scheme for Font Support

The scoring scheme is designed such that fewer points are deducted in case of lack of support of non-critical characters. Fonts that score higher provide better support for a particular language.

Font style and readability was analyzed with respect to different aspects like line height, curves and joins, kerning, hinting and other script specific features. User feedback was also taken into consideration for this purpose. Rendering speed for web fonts can critically affect the usability of web content. Font file size was used as an approximate measure for comparing font rendering speed. Licensing is another important aspect to consider while analyzing fonts. Fonts available under open licenses can be adjusted as per requirements by users and can be used in a wider variety of contexts.

Font embedding is becoming a critical aspect for enabling languages on the web. This is because computer systems are not usually shipped with Urdu fonts and a normal user may not know how to install such a font on his or her machine. Font embedding provides a convenient solution, where fonts are included in the content of the website and web pages are properly displayed even if the font is not installed on the user machine, as they are downloaded along with the webpage being accessed. Therefore, font embedding is also taken into account. The embedding analysis is carried out for different combinations of browsers and operating systems.

Nastalique is the conventional writing style for Urdu, though Naskh style has also been in use (Wali et al. 2006), especially in case of typewriters. Therefore, five available Nastalique fonts and one popular Naskh font are analyzed.

Web Fonts Assessment for Urdu: Appendix B below gives a summary of the results. The character support percentage is calculated using the scheme in Table 2 divided by the maximum score possible. The font samples are shown in selected form in order to show box height which has a significant impact on font readability (Urdu Web Interface Guidelines & Conventions Project, 2008a and 2008b). Nafees Nastalique and Nafees Web Naskh are found to be the most suitable fonts for use on the web for Urdu.

3.4 Collation Sequence

Lexicographic sorting of textual data should also be supported if multilingual content is to be developed for the web. The collation sequence should be first linguistically determined and standardized at national level, and then incorporated technically. Unicode (2010a) provides specific mechanisms for implementing the collation sequence, using collation elements for the characters in a language (Davis 2010). A more detailed language based analysis is given by Hussain et al. (2008b).

Collation Sequence Assessment Methodology: Collation sequence should be assessed at two levels. First the sequence of characters must be defined at a linguistic level. At the second level, collation elements should be defined to realize the sequence computationally. Finally these should be made part of national and international standard, e.g. Common Locale Data Repository.

Collation Sequence Assessment for Urdu: Urdu character sequence was recently finalized by the National Language Authority based on earlier work by Urdu Dictionary Board and given in Figure 3 below. Corresponding collation elements have also been suggested by Hussain et al. 2008b).

ا آ ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ

ڈھ ذ ر رھ ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ ف ق

ک کھ گ گھ ل لھ م مھ ن نھ ں ںھ و وھ ہ ۃ ء ی یھ ے

Figure 3: Urdu Collation Sequence

3.5 Locale

Among other details, formats for date, time, currency and measurement units and other similar elements are categorized as locale elements and they are used to display local language information online.

Locale Assessment Methodology: A comprehensive study of old and current publications is conducted to identify format conventions for dates, currency and measurement units.

Locale Assessment Results for Urdu: Two date formats, two time formats, a currency format, and measurement units as per the standard metric system are identified for use within the web content based on the use in correspondence conventionally. However these need to be standardized at national level. The selected date and time formats are shown in Figure 3.

۲۰ مارچ ۱۹۶۰ء ۰۵-۰۸-۲۰۱۰ء	۵ بجے ۸ بج کر ۲۳ منٹ

Figure 3: Selected date (left) and time (right) formats

3.6 Interface Terminology

Standardized translation of key web terms must be performed based on community feedback, in order to ensure a consistent web interface for users across applications and web pages.

Interface Terminology Assessment Methodology: At least the common terms being used online should be identified through a survey of websites and common desktop applications, translated based on existing standards and community recommendations (through open source translations).

Interface Terminology Assessment for Urdu: 1000 terms identified through the mechanism proposed above have been translated. Where possible, they are translated as recommended by the National Language Authority. Where the translations are not found within these recommendations, other translations and transliterations available through the online community are consulted and finalized through consensus of a team including linguists and technologists.

3.7 HTML Tags

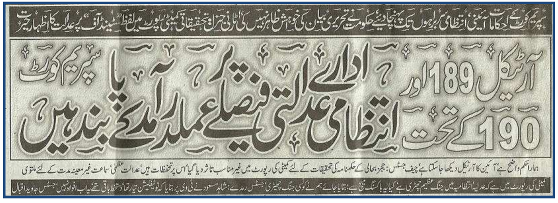
The standard HTML tag set has been designed in the context of Latin text formatting conventions. Publishing conventions of a language need to be identified and HTML tags need to be evaluated in the context.

HTML Tags Assessment Methodology: The tag analysis for Urdu was preceded by a survey of historical and current publications in order to identify conventions for elements such as heading styles, list styles, table styles, formatting styles for poetry, emphasis styles, footnote styles and other elements.

The HTML tags were then categorized as relevant (tags which can be directly implemented in their existing form), not-relevant (not applicable to the target language), adjusted (tags that need to be adjusted for use by the target language), proposed (additional tags for features that are required but are not available), enhanced (tags that require some enhancement in functionality to provide complete support).

HTML Tags Assessment for Urdu: A tag analysis is undertaken indicating only 74% of the HTML tags are directly usable for Urdu, with 4% not relevant, 15% need adjustment, 4% need enhancement and 3% new tags need to be defined.

As an example emphasis (em) tag needs to be adjusted for Urdu language. The normal behavior for this tag is to italicize text. However, italicization is not possible in Urdu calligraphy. However, emphasis is done in Urdu by writing text in white on black background, as in Figure 4.



سپریم کورٹ کے احکامات آئینی انتظامی سربراہوں تک پہنچا دیئے حکومت نے تحریری بیان کی خواہش ظاہر نہیں کی اٹارنی جنرل تحقیقاتی کمیٹی رپورٹ میں لفظ سٹینڈ آف پر عدالت کا اظہار حیرت

آرٹیکل 189 اور 190 کے تحت انتظامی ادارے عدالتی فیصلے پر عملدرآمد کے پابند ہیں سپریم کورٹ

ہمارا حکم واضح ہے آئین کا آرٹیکل دیکھا جا سکتا ہے چیف جسٹس: ججز بحالی کے حکمنامہ کی تحقیقات کے لئے کمیٹی کی رپورٹ میں غیر مناسب تاثر دیا گیا اس پر تحفظات ہیں عدالت عظمیٰ سماعت غیر معینہ مدت کے لئے ملتوی

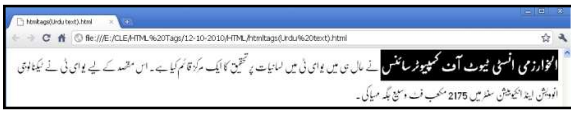
کمیٹی کی رپورٹ میں ہے کہ عدلیہ انتظامیہ میں جنگ عظیم چھڑی ہے کیا یہ باکسنگ میچ ہے: بتایا جائے ہم نے کونسی جنگ چھیڑی جسٹس رمدے: شاہد مسعود نے ٹی وی پر بتایا کہ نوٹیفکیشن تیار تھا دستخط باقی تھے یہ افواہ نہیں: جسٹس جاوید اقبال

Figure 4: Sample from an Urdu newspaper showing emphasized text.

This adjustment of the em tag is defined using style sheet, as given below:

```
/* empahsis tag */
em{
font-style:normal; /* font style
set to normal for text */
 background:#000000; /* back-
ground color set to black */
color:#FFFFFF; /* text color set
to white */
font-weight:600;
font-size:24px;
text-decoration:none;
}
```

This effect of adjusting the em tag is shown in Figure 5, where the text is displayed in white with a black background.


htmltags(Urdu text).html
file:///E:/CLE/HTML%20Tags/12-10-2010/HTML/htmltags(Urdu%20text).html

الخوارزمی انسٹی ٹیوٹ آف کمپیوٹر سائنس نے حال ہی میں یو ای ٹی میں لسانیات پر تحقیق کا ایک مرکز قائم کیا ہے۔ اس مقصد کے لیے یو ای ٹی نے ٹیکنالوجی
انوویشن اینڈ انکیوبیشن سنٹر میں 2175 مکعب فٹ وسیع جگہ مہیا کی۔

Figure 5: Adjusted em tag applied to Urdu text

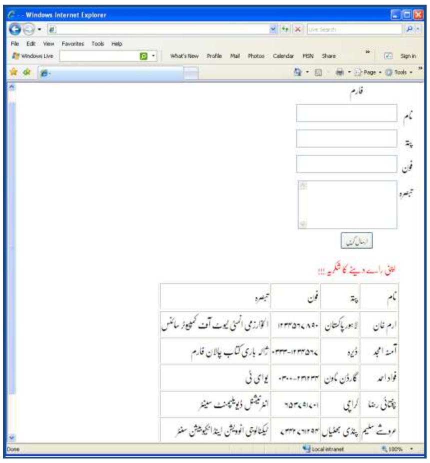
An additional tag is needed to support localized ordered lists for Urdu language. This exists at application level in some cases, however there is no support for it in the HTML standard.

Overall, the list of non-relevant tags for Urdu included the <i>, <rt>, <rp> and <ruby>. The list of tags that were adjusted for Urdu includes: <a>, <cite>, <ins>, <pre>, <textarea>, <address>, <code>, <kbd>, <samp>, <b>, <em>, <ol>, <select>, <button>, <input>, <option>, <strong>.

3.8 Web Technologies

Web technologies, in particular client, server and database technologies need to be tested to ensure that proper language support is available.

Web Technologies Assessment Methodology: The analysis for Urdu included server side scripting and database connectivity, in particular PHP with MySQL; ASP.net with Microsoft SQL Server. Display of local language strings in program editors and database fields are checked to ensure proper support, in addition to client end display. In addition, a form is designed to input and display information for testing purposes, shown in Figure 6. The default setting for fonts and dimensions are changed to adjust for Nastalique style.


Windows Internet Explorer
Live Search
File Edit View Favorites Tools Help
Windows Live
What's New Profile Mail Photos Calendar MSN Share
Sign in
Page Tools
Done
Local intranet
100%

فارم

نام □
پتہ □
فون □
تبصرہ □

ارسال کریں

آپکی رائے دینے کا شکریہ!!!

نام	پتہ	فون	تبصرہ
ارم خان	لاہور پاکستان	۱۲۳۴۵۶۷۸۹۰	الخوارزمی انسٹی ٹیوٹ آف کمپیوٹر سائنس
آمنہ امجد	ڈیرہ	۰۳۳۳-۱۲۳۴۵۶۷	ثار باری کتاب چالان فارم
فواد احمد	گارڈن ٹاؤن	۰۳۰۰-۲۳۱۲۴۴	یو ای ٹی
چغتائی رضا	کراچی	۶۵۴۳۲۱۰۱	انٹرنیشنل ڈیولپمنٹ سینٹر
عروشے سلیم	پنڈی بھٹیاں	۰۳۴۲۶۱۲۹۳	ٹیکنالوجی انوویشن اینڈ انکیوبیشن سنٹر

Figure 6: Form for web technology support assessment.

Web Technologies Assessment for Urdu: PHP/MySQL and ASP.net/Microsoft SQL Server were found to be generally supportive of Urdu and can be used for Urdu web applications. In both cases characters are also displayed properly within the database used.

3.9 Applications

Applications Assessment Methodology: A survey on advanced application support was also conducted to identify available applications which facilitate the uses online. This included typing tutors, spell checkers, online and desktop dictionaries, transliterators, machine translation systems, text to speech systems, automatic speech recognition systems and optical character recognition systems.

Advanced Applications Assessment for Urdu: No practically usable typing tutors are available for Urdu. Three pluggable spellcheckers are identified. Two in particular can be used in conjunction with Mozilla Firefox and OpenOffice.org. In addition, there are several useful online dictionaries and one desktop dictionary for Urdu. There are two transliteration utilities from ASCII to Urdu, both of which give reasonably robust transliteration results. Finally, there are two translation systems, one by Google and other by Ministry to IT; however neither is practically usable. Much more work needs to be done in this area.

4 Discussion and Conclusion

The analysis conducted shows that for Urdu language there are still some gaps in support.

Firstly, even though two fonts have been recommended through this study, the development of more optimal fonts is still needed. Secondly, some additional HTML support is needed. Thirdly, a lot more work is needed to provide adequate advanced application support.

Some of these gaps can be addressed through minor work-arounds, for example the adjustment of HTML tags through CSS. For other issues, updates are required in the standard, for example, support for localized ordered lists within HTML.

It is recommended that this analysis framework be used for other languages to assess support for online publishing and to identify and address gaps. These analyses can assist global efforts to provide support to create a truly multilingual web.

Acknowledgments

This work has been funded by the Punjab Information Technology Board of the Government of the Punjab, Pakistan.

References

Constable, Peter and Nelson, Jan Anders. 2010. *Bridging Languages, Cultures and Technologies*. The Multilingual Web – Where are we (Workshop), Madrid, Spain.

Davis, Mark. 2009. “Unicode Bidirectional Algorithm,” The Unicode Consortium. accessed from http://www.unicode.org/reports/tr9/ on 22nd Sept. 2010.

Davis, Mark and Whistler, Ken. 2010. Unicode Collation Algorithm 6.0.0. Unicode Consortium. Accessed from http://unicode.org/reports/tr10/.

Froumentin, Max. 2010. *The Remaining 5 Billion: Why is Most of the World’s Population not Online and What Mobile Phones Can Do About It*. The Multilingual Web – Where are we (Workshop), Madrid, Spain.

Hussain, Sarmad and Mohan, Ram. 2008a. Localization in Asia Pacific. In Digital Review of Asia Pacific 2007-2008. Orbicom and the International Development Research Center 2008.

Hussain, Sarmad and Durrani, Nadir. 2008b. A Study on Collation of Languages from Developing Asia. PAN Localization Project, International Development Research Center, Canada.

Hussain, Sarmad, Durrani, Nadir and Gul, Sana. 2005. PAN Localization Survey of Language Computing in Asia 2005. PAN Localization Project, International Development Research Center.

Hussain, Sarmad. 2003. www.LICT4D.asia/Fonts/ Nafees_Nastalique, in the Proceedings of 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore.

Hussain, Sarmad and Zia, Khaver. 2002. “Proposal to Add Marks and Digits in Arabic Code Block (for Urdu)”, in the Proceedings of 42nd Meeting of ISO/IEC JTC1/SC2/WG2, Dublin, Ireland.

Hussain, Sarmad and Afzal, Muhammad. 2001. Urdu Computing Standards: UZT 1.01, in Proceedings of the IEEE International Multi-Topic Conference, 2001, Lahore.

Internet World Stats 2010, www.internetworldstats.com

Ishida, Richard. 2010. *The Multilingual Web: Latest Developments at the W3C/IETF*. Workshop on The Multilingual Web – Where are we, Madrid, Spain.

Klensin, John. 2010. RFC 5891: Internationalized Domain Names in Applications (IDNA): Protocol. Internet Engineering Task Force. Accessed from http://tools.ietf.org/html/rfc5891.

Lata, Sawaran. 2010. *Challenges of Multilingual Web in India: Technology Development and Standardization Perspective*. The Multilingual Web – Where are we (Workshop), Madrid, Spain.

Lewis, M. Paul (ed.). 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: www.ethnologue.com

Lieske, Christian and Sasaki, Felix. 2010. *WC3 Internationalization Tag Set*. The Multilingual Web – Where are we (Workshop), Madrid, Spain.

PAN Cambodia 2007. Gap Analysis of HTML for Khmer, http://panl10n.net/english/Outputs%20Phase%202/CCs/Cambodia/MoEYS/Papers/2007/0701/ HTML_Standard_for_EnglishAndKhmer.pdf. PAN Localization Project, Cambodia.

Unicode 2010a. Unicode 5.0. 5th Edition. Addison-Wesley Professional, USA.

Urdu Web Interface Guidelines & Conventions Project. 2008a. “Urdu Web Font Evaluation Criteria”, University of Management and Technology. Unpublished report.

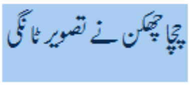
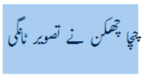
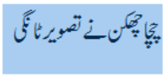
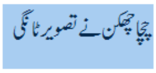
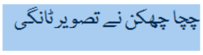
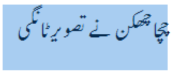
Urdu Web Interface Guidelines & Conventions Project. 2008b. “Usability Testing Report of Urdu Web Fonts”, University of Management and Technology. Unpublished report.

Wali, Amir and Hussain, Sarmad. 2006. *Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation.* In the Proceedings of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS2E 06).

Appendix A. Urdu Character Set within the Unicode Arabic Code Block

Unicode	Char
0600	؀
0601	؁
0602	؂
0603	؃
0604	
0605	
0606	□
0607	□
0608	□
0609	□
060A	□
060B	؋
060C	،
060D	؍
060E	؎
060F	؏
0610	◌ؐ
0611	◌ؑ
0612	◌ؒ
0613	◌ؓ
0614	◌ؔ
0615	◌ؕ
0616	□
0617	□
0618	□
0619	□
061A	□
061B	؛
061C	
061D	
061E	؞
061F	؟
0620	
0621	ء
0622	آ
0623	أ
0624	ؤ
0625	إ
0626	ئ
0627	ا
0628	ب
0629	ة
062A	ت
062B	ث
062C	ج
062D	ح
062E	خ
062F	د
0630	ذ
0631	ر
0632	ز
0633	س
0634	ش
0635	ص
0636	ض
0637	ط
0638	ظ
0639	ع
063A	غ
063B	□
063C	□
063D	□
063E	□
063F	□
0640	ـ
0641	ف
0642	ق
0643	ك V
0644	ل
0645	م
0646	ن
0647	ه V
0648	و
0649	ى V
064A	ي V
064B	◌ً
064C	◌ٌ
064D	◌ٍ
064E	◌َ
064F	◌ُ
0650	◌ِ
0651	◌ّ
0652	◌ْ
0653	◌ٓ
0654	◌ٔ
0655	◌ٕ
0656	◌ٖ
0657	◌ٗ
0658	◌٘
0659	◌ٙ
065A	◌ٚ
065B	◌ٛ
065C	◌ٜ
065D	◌ٝ
065E	◌ٞ
065F	
0660	٠ V
0661	١ V
0662	٢ V
0663	٣ V
0664	٤ V
0665	٥ V
0666	٦ V
0667	٧ V
0668	٨ V
0669	٩ V
066A	٪
066B	٫
066C	٬
066D	٭
066E	ٮ
066F	ٯ
0670	◌ٰ
0671	ٱ
0672	ٲ
0673	ٳ
0674	ٴ
0675	ٵ
0676	ٶ
0677	ٷ
0678	ٸ
0679	ٹ
067A	ٺ
067B	ٻ
067C	ټ
067D	ٽ
067E	پ
067F	ٿ
0680	ڀ
0681	ځ
0682	ڂ
0683	ڃ
0684	ڄ
0685	څ
0686	چ
0687	ڇ
0688	ڈ
0689	ډ
068A	ڊ
068B	ڋ
068C	ڌ
068D	ڍ
068E	ڎ
068F	ڏ
0690	ڐ
0691	ڑ
0692	ڒ
0693	ړ
0694	ڔ
0695	ڕ
0696	ږ
0697	ڗ
0698	ژ
0699	ڙ
069A	ښ
069B	ڛ
069C	ڜ
069D	ڝ
069E	ڞ
069F	ڟ
06A0	ڠ
06A1	ڡ
06A2	ڢ
06A3	ڣ
06A4	ڤ
06A5	ڥ
06A6	ڦ
06A7	ڧ
06A8	ڨ
06A9	ک
06AA	ڪ
06AB	ګ
06AC	ڬ
06AD	ڭ
06AE	ڮ
06AF	گ
06B0	ڰ
06B1	ڱ
06B2	ڲ
06B3	ڳ
06B4	ڴ
06B5	ڵ
06B6	ڶ
06B7	ڷ
06B8	ڸ
06B9	ڹ
06BA	ں
06BB	ڻ
06BC	ڼ
06BD	ڽ
06BE	ھ
06BF	ڿ
06C0	ۀ
06C1	ہ
06C2	ۂ
06C3	ۃ
06C4	ۄ
06C5	ۅ
06C6	ۆ
06C7	ۇ
06C8	ۈ
06C9	ۉ
06CA	ۊ
06CB	ۋ
06CC	ی
06CD	ۍ
06CE	ێ
06CF	ۏ
06D0	ې
06D1	ۑ
06D2	ے
06D3	ۓ
06D4	۔
06D5	ە
06D6	◌ۖ
06D7	◌ۗ
06D8	◌ۘ
06D9	◌ۙ
06DA	◌ۚ
06DB	◌ۛ
06DC	◌ۜ
06DD	۝
06DE	۞
06DF	◌۟
06E0	◌۠
06E1	◌ۡ
06E2	◌ۢ
06E3	◌ۣ
06E4	◌ۤ
06E5	ۥ
06E6	ۦ
06E7	◌ۧ
06E8	◌ۨ
06E9	۩
06EA	◌۪
06EB	◌۫
06EC	◌۬
06ED	◌ۭ
06EE	ۮ
06EF	ۯ
06F0	۰
06F1	۱
06F2	۲
06F3	۳
06F4	۴
06F5	۵
06F6	۶
06F7	۷
06F8	۸
06F9	۹
06FA	ۺ
06FB	ۻ
06FC	ۼ
06FD	۽
06FE	۾
06FF	ۿ

Appendix B. Font Analysis Results for Urdu

Font	Sample	Character support	Style and readability	Font file size	Embedding	License
Nafees Nastalique	 چچا چھکن نے تصویر ٹانگی	96%	High	388KB	Embeddable with some minor issues	Open
Alvi Nastalique	 چچا چھکن نے تصویر ٹانگی	98%	High	9MB	Embeddable but impractical	Free, for personal use only
Fajer Noori Nastalique	 چچا چھکن نے تصویر ٹانگی	93%	Low	255KB	Embeddable	Information not available
Jameel Noori Nastalique	 چچا چھکن نے تصویر ٹانگی	98%	High	13MB	Embeddable but impractical	Free for use
Nafees Web Naskh	 چچا چھکن نے تصویر ٹانگی	99%	Low	124KB	Embeddable	Open
Pak Nastalique	 چچا چھکن نے تصویر ٹانگی	96%	Low	167KB	Embeddable	Free for use