

# Multi Language Text Editor for Burushaski and Urdu through Unicode

Irfan Qadir Baig, Muhammad Sharif, and Aman Ullah Khan

**Abstract**—This paper introduces an isolated and unique ancient language Burushaski, spoken in Hunza, Nagar, Yasin and parts of Gilgit in the Northern Areas of Pakistan.

It explains the working mechanism of Multi Language Text Editor for Urdu and Burushaski. It is developed under the use of ISO/IEC 10646 Unicode standards for Urdu and Burushaski open-type fonts. It gives an ample opportunity to this regional ancient language to have a modern Information technology for its promotion and preservation. The main objective of this research paper is to help preserve the heritage of such rare languages and give smart way of automation. It also facilitates to those who are interested in undertaking research on Burushaski or keen to trace fonatic relationship between the national Urdu language and Burushaski.

Since this editor covers both Burushaski and Urdu so it can play an important role to introduce Burusho linguistic culture to the world at large.

Precisely, as a result of this research paper, Burushaski publication through IT means would be possible.

**Keywords**—Burushaski, Bri Naqsh, Unicode, Burusho, Hunza, Meshaski.

## I. INTRODUCTION

UNICODE is an encoding system that provides a unique number for every character and it does not matter what the platform is and it does not matter what the language is [1].

Before Unicode there were different encoding systems to allocate number to those characters. Due to limitations of those encoding systems, people understand that they need a powerful encoding system that could address the needs of all languages. Although the Unicode still has limitations but comparatively is better than other local encoding systems used for this purpose.

### A. The Unicode Standard

Unicode standard is defined by an organisation known as Unicode Consortium. In 1991 they proposed world wide acceptable character encoding system [1].

ISO/IEC 10646 standard accepted by International Standard Organization (ISO) which defines International information processing standards; these standards are set though votes of all members of the consortium [1].

Authors are with Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Punjab, Pakistan (e-mail: iqbaig@yahoo.com, muhammadsharifmalik@yahoo.com, auk\_pk@yahoo.com).

In 1991 Unicode organization released its first Unicode standard. This 16 bit ISO Unicode standard set encoding specifications; enabling 65,535 unique characters of various languages of the world. These uniquely encoded characters enable us to address large languages like Chinese, Korean.

### B. Unicode for Perso-Arabic Script

Urdu uses the Arabic script with extensions. A number of extensions are based on those developed for Persian. Burushaski share same extensions plus some unique alphabets. The Urdu alphabets falls in Arabic range from (0600-06FF) Hex. Burushaski shares the same script so some of its alphabets fall in same range. Fig. 2 shows the complete Urdu Burushaski alphabets.

### C. Open Font Type

The algorithm is used for Unicode Character representation for Urdu that is available in open-type fonts or those fonts that build with ISO/IEC 10646 Unicode support. Open type Bki Naqsh fonts are designed for Burushaski. Bki Naqsh is little different than the Urdu Nafiz Naqsh font available. There are also open fonts that are available for other languages. Some of them are

1. Nafiz Naqsh
2. Time Roman
3. Bki Naqsh.

### D. How to Create Personal Unicode Based Characters?

Font creating tools can be used to create personal true type fonts. High-Logic's Font Creator 4.5 is used to create Burushaski open type fonts. It follows Unicode standard. Through this way authors accommodated unique Burusho characters.

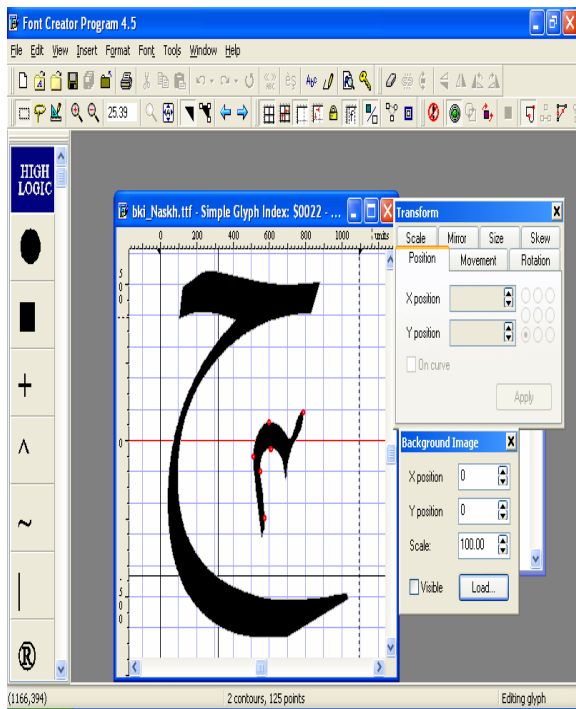


Fig. 1 Open Type Burushaski Font Creation

## II. BURUSHO ORIGIN

Although it is difficult to say about the origin of Burusho (Burushaski speakers) but some researchers believe that they were settled originally in Hunza, Yasin, Gilgit and Baltistan of Northern Areas, Pakistan. Gilgit was the capital of Burusho Kingdom called Burushal. Due to foreign invasions from Tibet and Indus valleys, Burusho scattered into three valleys, Hunza, Yasin and Nager.

### A. Burushaski Language

About 150,000 people speak burushaski in three deferent dialects Meshaski in Hunza, Khajona in Nagar and Werchikwar in Yasin in Karakurum region, Northern part of Pakistan. Burushaski language is not obviously related to any of the surrounding language; the Indic language of Pakistan, Tibetan Language of China and Northren Kashmir [2]

This isolated language always attracts researchers toward it, but still no relation found with other contemporary or dead languages in the history of mankind. Rulin still classified Burushaski as an isolated language. Its generic affiliation remains complete mystery. Some researchers are agreed to give it a newly branch [3].

### B. Burushaski Alphabets

Linguists have described Burushaski as spoken language because of its limited written scripts.

But from last 50 years its native speakers did a valuable work along foreigner researchers. In 1940's Dr. Prof. Nasir-al-din Hunzai selected Burushaski to express his Sufi Poetry. Prof. Hunzai is known as father of Burushaski in view of his devoted services for this language. He wrote hundreds of poems in his native language. He has collected about 60,000

Burushaski words through different means for first Burushaski German Dictionary which is in progress to be published. His first poem came in 1940. He gave first Perso-Arabic script to this language along Roman alphabets. Foreigners like D.L.R. Lorimar [4], Herman Bearge[5], and E.N. Teffue [6] worked on Roman Script for this language.

### C. Perso-Arabic Script

Perso-Arabic alphabets proposed for this language by Dr. Nasir-u-din Hunzai. Prof. Hunzai used thirty six (36) Urdu alphabets and seven extra alphabets used to pronounce this language correctly [6].

TABLE I  
 CHARACTER RENDERING ALGORITHM

Isolated	Final	Meddle	Initial
چ	چ	چ	چ
خ	خ	خ	خ
د	د	.	.
ذ	ذ	.	.
س	س	س	س
ص	ص	ص	ص
ش	ش	ش	ش

The Urdu alphabets falls in Arabic range from (0600-06FF) Hex. Fig. 2 shows the complete Urdu Burushaski alphabets.

ث	ث	ت	پ	ب	ا
خ	خ	ح	چ	چ	ج
ز	ر	د	ذ	ذ	د
ص	س	ش	س	ژ	ز
غ	ع	ظ	ط	ض	ض
ل	گی	شی	ک	قی	ف
ھ	ہ	و	ن	ن	م
		ہ	ی	ی	ء

Fig. 2 Complete Burushaski Alphabets

### D. Character Rendering

After applying bidirectional algorithm the order of character is called visual order. This is the order that they should appear on the screen. The Arabic joining algorithm determines which

character should be rendered. After that some ligatures may form.

Since Burushaski alphabets are developed from Arabic and Persian script so the letters joining mechanism is same. Letters joining algorithm defines the behaviour of each letter.

From the above special Burushaski characters few are also seen in other languages, like Sindhi, Pushto etc. Table II shows the list of such characters. The pronunciation of these alphabets could be different [8]. The list of alphabets also presents in other languages.

TABLE II  
 SPECIAL BURUSHASKI CHARACTERS ALSO AVAILABLE IN OTHER LANGUAGES

Character	Language	Unicode
٤	Burushaski/ Sindhi	<b>068E</b>
٥	Burushaski/ Berber	<b>069E</b>
٦	Burushaski/ Pashto	<b>0685</b>
٧	Burushaski/ Persian..	<b>0698</b>

As Burushaski is an isolated and different language it has some alphabets that are unique in world languages and give different sounds.

TABLE III  
 SPECIAL UNIQUE BURUSHASKI CHARACTERS

Character	Description
٤	Arabic HAY with four dots = Indo Arabic digit four below
٥	Arabic letter SEEN, with four dots = Indo Arabic digit four above.
٦	Perso-Arabic letter KAFF with three dots above.
٧	Arabic letter Alif maksura with shape like YEH with four dots below = Indo Arabic digit four.

#### E. Burushaski Number Symbols

Burushaski follows same number symbol used in Arabic language.

TABLE IV  
 BURUSHASKI DIGITS

Symbol	Unicode	Description
٠	06F0	Eastern Arabic Indic digit Zero
١	06F1	Eastern Arabic Indic digit one
٢	06F2	Eastern Arabic Indic digit Two
٣	06F3	Eastern Arabic Indic digit Three
٤	06F4	Eastern Arabic Indic digit Four
٥	06F5	Eastern Arabic Indic digit Five
٦	06F6	Eastern Arabic Indic digit Six
٧	06F7	Eastern Arabic Indic digit Seven
٨	06F8	Eastern Arabic Indic digit Eight
٩	06F9	Eastern Arabic Indic digit Nine

### III. MULTILANGUAGE TEXT EDITOR

Unicode has made possible to develop Multilanguage text editor for Burushaski and Urdu. The basic purpose of this editor is to automate Burushaski language and preserve this ancient heritage. Urdu users can also use it to compose their text, as Burushaski has all Urdu alphabets and properties, plus some extra alphabets and extensions. Multilanguage text editor for Burushaski and Urdu created with the help of Unicode and Unicode based open type-fonts. Since many alphabets are totally unique in Burushaski, so Bki Naqsh fonts are created for this language.

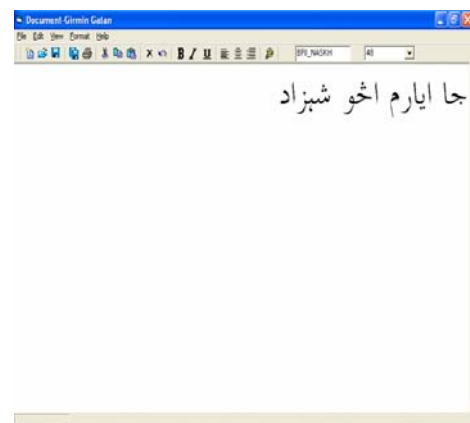


Fig. 3 Burushaski Editor

The editor gives facility to compose text, save and print in standard form.

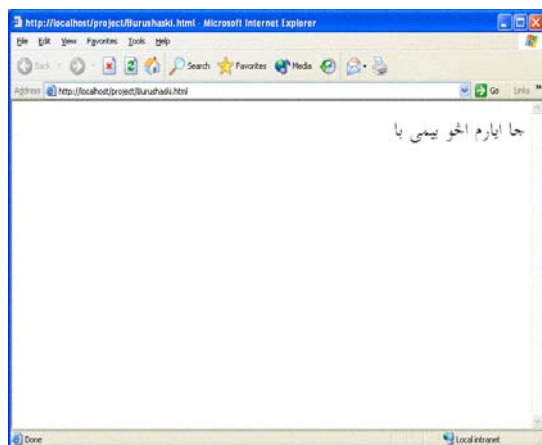


Fig. 4 Burushaski Text in Web Browser

- [4] Loimer, D.L.R., The Burushaski Language – Vol.1,2,3, Institute for Sammenligende Kulturforskning, Oslo, 1935,464 pp.
- [5] Berger, Hermann, Das Yasni Burushaski (Warchikwar) 1974.
- [6] Tiffou, Etienne, Jurgen Pesot, Content du Yasin- Introduction au bourouchaski du Yasin, Etudes bourouchski 1, paris, 1989.
- [7] Dr. Allama Nasir al-din Hunzai, Diwan-e-Nasiri, Published by Burushaski Research Academy Karachi Pakistan.
- [8] [www.unicode.org/charts/](http://www.unicode.org/charts/)
- [9] <http://en.wikipedia.org/wiki/Unicode>

Unicode made possible to save Burushaski and Urdu text in HTML form and run on any Unicode supported Internet browser. With IE5 only need to enable the UTF- 8 option from view menu bar.

#### IV. CONCLUSION

The development of Multilanguage Text Editor for Burushaski and Urdu made possible by Unicode. The system made possible to automate and preserve the heritage of an ancient language Brushaski. It is believed that it is still in its original form. The system provides all text editor related operations of both languages.

#### V. FUTURE WORK

Since Burushaski is an isolated and unique language and mostly the research work done by western researchers so they proposed Roman Script to this language. Local researchers like Prof. Nasir Hunzai are agreed that Roman Alphabets pronounce this language more correctly than the Arabic Script. But still it needs to give standardise Roman alphabets. The Roman Alphabets will enhance the capabilities of this software along Arabic Script. It will provide choice for users.

#### ACKNOWLEDGMENT

Authors are thankful to Dr. Prof. Nasir Hunzai for his valuable work in preservation of world's unique ancient language who gave it Arabic Script and Roman Alphabets. We are also thankful to all foreign researchers who spent their precious time in such difficult clement and done valuable work for promotion and preservation of this language.

#### REFERENCES

- [1] [www.unicode.org/unicode/standard/WhatIsUnicode.html](http://www.unicode.org/unicode/standard/WhatIsUnicode.html).
- [2] Burushaski an Extraordinary language in the Karakorum region by Dick Grune August 17, 1998.
- [3] Ruhlen,Marritt, A Guid to the world's Languages: Volume 1: Edward Arnold, London,1987;1991, 463pp.