ISSN: 2223-9553
Volume 1, Issue 1, July 2011

SURVEY OF COMPUTATIONAL SUPPORT FOR SHAHMUKHI SCRIPT OF PUNJABI LANGUAGE

Saira Javaid
PUCIT, University of the Punjab, Lahore
PAKISTAN
ajee_4me@hotmail.com

Hira Sattar
PUCIT, University of the Punjab, Lahore
PAKISTAN
hira.sattar88@gmail.com

Aasim Ali
PUCIT, University of the Punjab, Lahore
PAKISTAN
aasim.ali@pucit.edu.pk

M. G. Abbas Malik
GETALP – LIG, University of Grenoble
FRANCE
abbas.malik@imag.fr

ABSTRACT

This paper is a survey of Punjabi language to explore the computational facilities available for orthography of this language. The two scripts used for writing Punjabi are Shahmukhi (a derivation of the Perso-Arabic script) and "Gurmukhi". Our main focus was on Shahmukhi, mainly used in Pakistan, to identify the support by Unicode for its characters.

Keywords: Punjabi, Unicode Support for Punjabi, Punjabi Language Processing, Computational Support for Punjabi.

INTRODUCTION

Unicode is a single character set developed to support several languages of the world. The purpose of developing this huge set was to provide a universal set that contains other such existing sets earlier. Unicode serves as a single medium of interpreting the data of several languages across the globe. Single document that contains text of multiple languages need not to have multiple character sets; only one big set of all characters of written languages may serve the purpose. If that set is understandable to every operating environment uniformly then there is no need to attach it along with the document while transmitting data from one platform to another. Some less known languages may also be incorporated with addition of any language specific characters and using the majority characters from the neighboring languages of the same scripting family.

Although Punjabi is not a less known language on the globe but for the purpose of its write ability, there is a need of availability of all its characters in the Unicode. This study investigates that to what extent this facility is available for Punjabi language.

In this paper section 2 defines benefits of Unicode, section, section 3 tells about "Shahmukhi" and its characteristics, section 4 defines about RNOON and RLAM and their problems, then section 5 shows Unicode mapping of consonants, section 6 lists Vowel and their Unicode, section 7 is about Numerical values, other Diacritical marks and Symbols, in section 8 we have conclusion.

BENEFITS OF UNICODE

Unicode has following advantages:

(1) The text of any language can be represented in Unicode. It also provides portability.

ISSN: 2223-9553
Volume 1, Issue 1, July 2011

(2) Unicode allows to use multilingual text using any or all the languages of desire.

(3) Unicode increases support and availability of software for local languages because of their limited scope. Otherwise such support is not available.

(4) Unicode really make it easier to globalize IT and provide advantages to all the people and languages.

Pakistan is a country with at least six major languages and 58 minor ones (Malik, 2005), as shown in Table 1 given below.

Table 1: Six major languages of Pakistan (taken from (Malik, 2005))

Language	Percent of Speakers	Number of Speakers
Punjabi	44.15	66,225,000
Pashto	15.42	23,13,000
Sindhi	14.10	21,150,000
Siraki	10.53	15,795,000
Urdu	7.57	11,355,000
Balochi	3.57	5,355,000
Others	4.66	6,990,000

The above table shows the importance of Punjabi. It is on top of all the languages spoken in Pakistan, on the basis of number of its speakers. Punjabi is also spoken in America, Canada, and Europe (as per Ethnologue, see References). There are two scripts which are used to write Punjabi: (1) Shahmukhi, a derivation of the Perso-Arabic script; and (2) Gurmukhi, a derivation of ancient Indic scripts like LANDA (script of north-west), SHARDA (script of Kashimir) and TAKRI (script of western Himalaya) (Malik, 2006). The focus of this study remained on Shahmukhi script.

CHARACTERISTICS OF SHAHMUKHI

. Shahmukhi derives its character set from Persian/Arabic scripts and came into writing nearly about 10th and 11th century after the establishment of the Mughal Empire in the Indian Subcontinent(Malik, 2005).

Shahmukhi character set is a super set of Arabic/Persian alphabets. It contains 43 characters and the 15 diacritical marks. Unlike English, the Characters do not have upper and lower case. Below Table shows the Punjabi characters set. (Malik, 2005)

ISSN: 2223-9553
Volume 1, Issue 1, July 2011

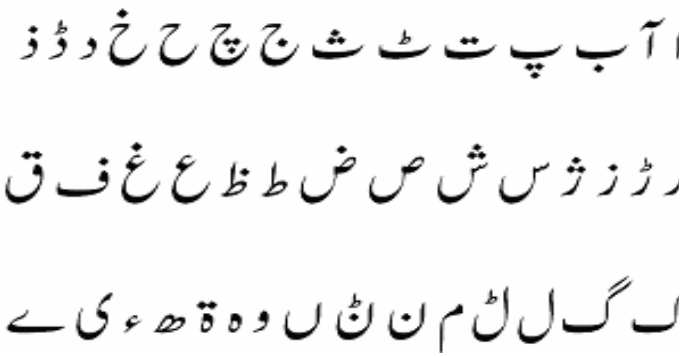

Figure 1: Character set of Punjabi (Shahmukhi) (taken from (Malik, 2005))

Shahmukhi is a right to left script and shapes assumed by characters in a word are context sensitive and the shape of a character depends on the position of the character, i.e., at the beginning, in the middle or at the end of a word (Malik 2005; Zia 1999). Thus a character may have initial, medial and final shapes depending on its context in the word and the fourth one is the standalone shape of the character.

Figure 2: Conext sensitive shapes of BEY (taken from (Malik, 2005))

The above is true for all except eleven characters. Ten of them have only two shapes; the standalone and the terminating final shape (Malik, 2005). These ten characters are shown in below:

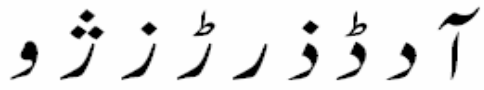

Figure 3: Characters having standalone and final shapes (taken from (Malik, 2005))

Hamza does not occur at the start of a word, but it comes in the start of a ligature. It has initial, middle and autonomous shapes (Malik, 2005) as illustrated in the Figure 4 given below.

Figure 4: Shapes of Hamza, (Circled, right to left) Independent, Initial and Middle shape (taken from (Malik, 2005))

0600h–06FFh and FB50h–FEFFh are the codes in Unicode that are associated with Arabic and languages written in its derivations like Punjabi and Urdu (as per Unicode Standards 2.5). Most of Shahmukhi characters are already in Unicode, but a few characters are missing.

RNOON AND RLAM

It is a general agreement that the sounds of RNOON and RLAM are present in the Punjabi language but they are written with different shapes. There are six different shapes for RNOON at the time (Malik, 2005). RLAM is rarely used in Punjabi but there are different opinions about its shape and its inclusion in Character set (Malik, 2005).

There is a convention that RNOON and RLAM never come in the beginning of the word and they are never found together in a single syllable. As mentioned above, the shape assumed by a character is context sensitive; thus both RNOON and RLAM have six different context sensitive shapes, i.e., medial, final and standalone shape in written words. They also assume their initial shapes when they come in the middle of the word and are proceeded by a character shown in Figure 3.

RNOON is also found in other languages written in the derivation of the Arabic script and it is assigned the Unicode values 06BB and FBA0 (both shapes on these codes are equivalent). RLAM is not defined in Unicode.

RNOON Problem

In the RNOON case, RNOON is present in all dialects of Punjabi. The problem with RNOON is to resolve its symbolic illustration. Six different shapes are suggested:

(1) Plain NOON ن – let the sound be resolute by context. This shape is the same as normal NOON and that leads to confusion.

(2) NOON with two vertical Dots replacing one Dot (used by Punjabi Adabi Board).

(3) NOON with TOAY mark as a replacement for dot ڻ (also use in Sindhi and accessible in Unicode). This type of shape is related to TTEY ٹ, particularly when it comes at the start & middle of the word.

(4) NOON with both Dot & TOAY mark above ﮟ.

(5) NOON with "kundi" like in Pashto ڼ. The distinction used in this shape is not used for any further character.

(6) NOON with small circle instead of dot ں. Small circle is a new mark not found in Punjabi.

RLAM Problem

RLAM may be added in the character set of Shahmukhi script. Although it is very rarely used but it should be integrated in the character set for Punjabi Shahmukhi as the sound of RLAM exists in the Punjabi language (Malik 2005).

In the table below, the missing Punjabi characters have been listed. Each character is given a symbol and proposed description. If these missing characters are given a place in Unicode standard, it would make Punjabi Shahmukhi compatible with Unicode and ISO/IEC 10646 (Malik, 2005)

Summary of Proceedings of Meeting (Punjabi baithak) of Codepage Subcommittee of Urdu and Regional Language Software Development Forum (URLSDF) of Ministry of IT held on 6th April, 2002 in Lahore

ISSN: 2223-9553
Volume 1, Issue 1, July 2011

Table 2: Unicode letters proposed for Shahmukhi (taken from (Malik, 2005))

Serial #	Symbol	Proposed Unicode	Description
1		063B	Arabic/Punjabi Shahmukhi RLAM
2		063C	Arabic/Punjabi Shahmukhi RNOON
3		0659	Arabic/Punjabi Shahmukhi SAKUN
4		0616	Arabic ligature Jalla Jalalouhou

CONSONANT MAPPING

This section lists consonant characters in Punjabi, along with their Unicode values. Consonants can be further subdivided into two groups:

Aspirated Consonants

Table 3 given below lists the Aspirated Consonants of Punjabi:

Table 3: Aspirated Consonants – Shahmukhi

Serial #	Character	Aspiration Mark	Aspirated Consonant	Unicode
1	ب	ھ	بھ	06BE+0628
2	پ	ھ	پھ	06BE+067E
3	ت	ھ	تھ	06BE+062A
4	ٹ	ھ	ٹھ	06BE+0679
5	ج	ھ	جھ	06BE+062C
6	چ	ھ	چھ	06BE+0686
7	د	ھ	دھ	06BE+062F
8	ڈ	ھ	ڈھ	06BE+0688
9	ڑ	ھ	ڑھ	06BE+0691
10	ک	ھ	کھ	06BE+06A9
11	گ	ھ	گھ	06BE+06AF

ISSN: 2223-9553
Volume 1, Issue 1, July 2011

Non-Aspirated Consonants

Table 4 given below lists the Non-Aspirated Consonants of Punjabi:

Table 4: Non-Aspirated Consonants – Shahmukhi

Serial #	Character	Unicode
1	ب	0628
2	پ	067E
3	ت	062A
4	ٹ	0679
5	ث	062B
6	ج	062C
7	چ	0686
8	ح	062D
9	خ	062E
10	د	062F
11	ڈ	0688
12	ذ	0630
13	ر	0631
14	ڑ	0691
15	ز	0632
16	ژ	0698
17	س	0633
18	ش	0634
19	ص	0635
20	ض	0636
21	ط	0637
22	ظ	0638
23	ع	0639
24	غ	063A
25	ف	0641
26	ق	0642
27	ک	06A9
28	گ	06AF
29	ل	0644
30	م	0645

31	ن	0646
32	و	0648
33	ہ	06C1
34	ی	06CC
35	ے	06D2
36	ڑ	0768
37	ں	----

VOWEL MAPPING

There are two types of vowel: Long vowels and Short vowels.

Long Vowels

Table 5 given below lists the short vowels of Punjabi:

Table 5: Long Vowels – Shahmukhi

Serial #	Vowel	Name	Unicode
1	ا	Alif	0627
2	آ	Alif Madda	0622
3	و	Vav	0648
4	ی	Ye Chhoti	06CC
5	ے	Ye Bari	06D2

The long vowel characters VAV (و) and YE (ی) may also behave like a consonant.

Short Vowels

Table 6 given below lists the short vowels of Punjabi:

Table 6: Short Vowels – Shahmukhi

Serial #	Vowel	Name	Unicode
1	◌َ	Zabar	064E
2	◌ُ	Pesh	064F
3	◌ِ	Zer	0650

ISSN: 2223-9553
Volume 1, Issue 1, July 2011

NUMERICAL VALUES AND OTHER

Table 7 given below lists the numeric symbols of Punjabi:

Table 7: Numeric Symbols – Shahmukhi

Numeric Value	Digit Symbol	Unicode
0	۰	06F0
1	۱	06F1
2	۲	06F2
3	۳	06F3
4	۴	06F4
5	۵	06F5
6	۶	06F6
7	۷	06F7
8	۸	06F8
9	۹	06F9

Table 8 given below lists the other symbols of Punjabi:

Table 8: Other Symbols – Shahmukhi

Serial #	Symbol	Unicode
1	ً	064B
2	ّ	0651
3	ٰ	0670
4	٪	066A
5	؛	061B
6	؟	061F
7	ـ	06D4
8	،	060C

CONCLUSION AND FUTURE WORK

The computational orthographic support for writing Punjabi language in Shahmukhi is already available in Unicode, except a character (RLAM), as per Unicode 5.2 [see References]. There is a need of spreading the awareness in the public to produce Punjabi contents. Government agencies may also support some projects for several (at least six major) languages of our nation.

During this study, some phonetic aspects of Punjabi also found to be worth investigating (for example, the voiceless behavior of aspirated sounds), but could not be included in this paper due to the scope of work.

ISSN: 2223-9553
Volume 1, Issue 1, July 2011

REFERENCES

Malik, M. G. A. (2005): *Towards a Unicode Compatible Punjabi Character Set*, 27th Internationalization and Unicode Conference, Berlin, Germany, April 2005

Malik, M. G. A. (2006): *Punjabi Machine Transliteration*, in the proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of ACL, pages 1137 – 1144, Sudney, Australia, July 2006

The Unicode Consortium. The Unicode Standard, Version 5.2.0, defined by: *The Unicode Standard, Version 5.2 (*http://www.unicode.org/versions/Unicode5.2.0/*)*

Ethnologue, Languages of the World; available at http://www.ethnologue.com/

Zia, K. (1999): *Towards Unicode Standard for Urdu*, in the Proceedings of 4th Symposium on Multilingual Information Processing (MLIT-4), Yangon, Myanmar, CICC, Japan.