TOWARDS A UNICODE COMPATIBLE PUNJABI CHARACTER SET

M. G. Abbas Malik
E-mail: abbas.malik@gmail.com

ABSTRACT

This paper discusses additions to the Arabic script for the representation of the Punjabi language. This adaptation of the Arabic script is called the "*Shahmukhi*" writing system which is one of the two writing systems of the *Punjabi* language, the other being *Gurmukhi*. To begin with, the representation of the Punjabi language in the two scripts is compared. Characteristics of different characters i.e. letters, special symbols, diacritical marks, digits etc, are analyzed.

Based on this analysis a new character set is proposed, which is capable of fully representing Punjabi in *Shahmukhi*. A comparison of this set with Unicode is undertaken to identify missing characters. Inclusion of these characters is proposed with a view to make *Shahmukhi* compatible with the Unicode standard.

Keywords

Codepage, Gurmukhi, Multilingual Processing, Punjabi, Shahmukhi, Standardization, Unicode

1. Introduction

The benefits from Information Technology (IT) revolution cannot be reaped unless the masses use it, which is not possible unless computing is possible in a language that is understood by masses. This realization has come to many nations already. According to Jones S et al. (1992), the German *Munchener Oberlandesgericht* court decision of 1985 restricts the delivery of computers if it does not accompany operating instructions in German. Similar measures have been taken in many of the European and Far-Eastern countries, to enforce their local/national languages [1]. Pakistan is a country with at least six major languages and 58 minor ones [2].

Table 1: Six major languages of Pakistan and number of their native speaker

Language	Percentage of Speakers	Number of Speakers
Punjabi	44.15	66,225,000
Pashto	15.42	23,130,000
Sindhi	14.10	21,150,000
Siraiki	10.53	15,795,000
Urdu	7.57	11,355,000
Balochi	3.57	5,355,000
Others	4.66	6,990,000
Source: Census 2001: Table 2.7. The population is assumed to be 150 million in 2003 as it was 132,352,000 in 1998 and the growth rate is 2.69 %.		

It is clear from the table above that Punjabi is most widely spoken language of Pakistan according to number of its native speakers.

Thus more than 66 millions people of Pakistan cannot enjoy the benefits of IT revolution unless standards for Punjabi – *Shahmukhi* are produced and then implemented. The need for standardization for computerized language processing cannot be overemphasized. The standardization has the same relation to computerized language processing as a standard keyboard has to typewriting [4].

The present paper first gives a brief comparison between two scripts used for Punjabi; this is followed by the characteristics of *Shahmukhi* script. It also discusses major hindrances and

problems in standardization of the character set. It then proposes a character set for *Shahmukhi* keeping in view the recommendations and findings of [5]. Finally, an analysis between the proposed character set and Unicode is done to make the set compatible with the Unicode and proposes the inclusion of native characters of Punjabi in Unicode.

2. Gurmukhi and Shahmukhi – a Brief Comparison

Muslims had started to influence the Indian Subcontinent since 9th century. *Shahmukhi* derives its character set from Persian/Arabic scripts. Its use to transcribe Punjabi commenced around 10th and 11th century after the *Mughal* conquest and establishment of vast empire in Indian Subcontinent [6]. It is a right to left script and the shape assumed by a character in a word is context sensitive and is used for Punjabi in *Pakistan*. In Unicode, Arabic and its associative languages like Punjabi, Urdu etc. have been allocated 1,200 code points (0600h – 06FFh, FB50h – FEFFh) and most Shahmukhi characters are already in Unicode, but a few characters are missing.

Gurmukhi derive its character set form Landa (old script of Indian sub-continent) and was standardized by *Guru Angad Dev* (second Sikh Guru, 1504 – 1552) in the 16th century and contained 35 consonants at that time. The word *Gurmukhi* literally means "from the mouth of Guru". The whole of the *Guru Granth Sahib* (Holy book of Sikhs) is written in *Gurmukhi*. Its alphabets are *Abugida*, as each consonant has an inherent vowel (a) that can be changed using vowel signs [7]. It is a left to right script and unlike Shahmukhi its characters do not assume different shapes and also do not have small and capital forms. An example sentence is give below:

ਪੰਜਾਬੀ ਮੇਰੀ ਮਾਣ ਜੋਗੀ ਮਾਂ ਬੋਲੀ ਏ.

Modern *Gurmukhi* has 41 consonants, 9 vowels symbols, 2 symbols for nasal sounds, 1 symbol that duplicates the sound of any consonant, 3 subjoined forms of the consonants Ra, Ha and Va and 1 post-base from of Ya [7]. In Unicode, *Gurmukhi* sub-range is from U+0A00 to U+0A7F. This provides 128 code points for Gurmukhi characters of which only 77 are currently used (Unicode 4.0.1). In addition, Danda and Double Danda are contained in the *Devanagari* sub-range at U+0964 and U+0965 respectively.

3. Characteristics of Shahmukhi

The distinguishing characteristics of the *Shahmukhi* are discussed for the benefit of the unacquainted reader. Punjabi is greatly influenced by *Arabic* and *Persian* languages. *Shahmukhi* derived its character set from Persian and Arabic and its character set is a super set of Arabic and Persian alphabets and contains 43 basic characters and 15 diacritical marks. Figure 1 shows the alphabets of *Shahmukhi*. Unlike English, the characters do not have upper and lower case.

ا آ ب پ ت ٹ ث ج چ ح خ د ڈ ذ

ر ڑ ز ژ س ش ص ض ط ظ ع غ ف ق

ک گ ل ڵ م ن ڻ ں و ہ ۃ ھ ء ی ے

Figure 1: Character set of Punjabi (Shahmukhi)

Further, the shape assumed by a character in a word is context sensitive i.e. the shape is different depending whether the position of the character is at the beginning, in the middle or at the end of the constituent word. This generates three shapes, the fourth being the independent shape of the character. Figure 2 gives these four shapes for a character, named *Bey*.

Figure 2: Context sensitive Shapes of Bey

To be precise, the above is true for all except eleven characters. Ten of these have only two shapes; the independent and the terminating shape, these characters are shown in Figure 3. These characters have independent and final shapes when they come at beginning and middle of a word respectively.

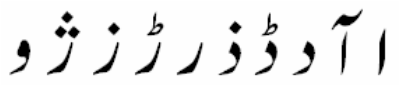


Figure 3: Characters having only Independent and Final Shape

Hamza never comes at the beginning of a word [4], but it comes in the beginning of a ligature. Also it attains the independent shape instead of the final shape when it comes at the end of the word. Owing to this, it has initial, middle and independent shapes. It is illustrated in figure 4.

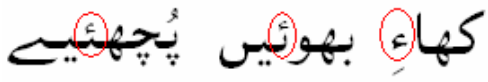


Figure 4: Shapes of Hamza, (Circled, right to left) Independent, Initial and Middle shape

Punjabi is traditionally written in *Nastaleeq*, a script rich in calligraphic content. Owing to complexities of rendering, the basic shapes identified above are unable to render the language in an acceptable form in *Nastaleeq* [4]. The characters of Punjabi also need diacritics to help in the proper pronunciation of the constituent word. The diacritics appear above or below a character to define a vowel or emphasize a particular sound. These diacritical marks are basis of the vowel system in Shahmukhi. There are a number of diacritics, the common ones being *Zabar*, *Zer*, and *Pesh* [4]. Figure 5 shows the character *Bey* marked with these diacritics.

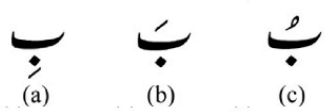


Figure 5: Bey with Diacritics

Figure 6 shows Punjabi text in *Nastaleeq* script with diacritics placed on the respective characters.

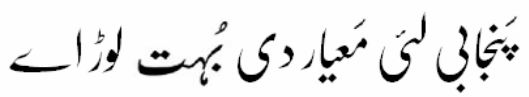


Figure 6: Punjabi (Shahmukhi) Text in Nastaleeq

Diacritics, though part of the language, are sparingly used. They are essential for removing ambiguities, natural language processing and speech synthesis [4]. Thus, (a) the multiple shapes of characters, (b) the complexities of the traditional script of Punjabi and (c) the existence of diacritics, are major factors that contributed to the difficulties in formulating a standard for Punjabi.

4. Character Set

In standardization of a language, first question, which has to be resolved, is the character set. Character set contains all characters, numerals, punctuation marks, symbols etc. required to write a language [1]. Different authors have quoted different number of characters for Punjabi. Mirza Maqbool Baig (1973) and Shehbaz Malik (1996) have proposed 52 and 54 characters respectively.

4.1. Points of Disagreement

Before embarking on the standardization of character set for Punjabi, it is necessary to discuss the controversial points and major hindrances in the process of standardization of character set for Punjabi.

4.1.1. RNOON and RLAM

It is agreed that RNOON and ARLAM are sounds of Punjabi [5] [10]. However there is a disagreement on the shape of this letter. Six different shapes are suggested for RNOON at the moment. RLAM is used very rare in Punjabi but there are different opinions about its shape and its inclusion in Character set [5].

4.1.2. Aspirated Characters

There are ten aspirates in Punjabi and are currently represented by the aspirated consonant + Do-Chashmey-Hay. There is a disagreement on inclusion of aspirates as separate characters in the character set. [5]

4.1.3. Numerals

There are two different opinions about the numerals of Punjabi. One is that traditional Urdu style numerals should be used and other is to use English numerals. [5]

4.2. Points of Agreement

It is agreed to include all the *diacritical, Punctuation* marks and *special symbols* of UZT 1.01 [11], except long mad, Hamza – Waw and Lam – Alef symbols in the standard character set for Punjabi [5].

5. Recommended Solutions

5.1. RNOON and RLAM

In Punjabi, RNOON and RLAM never come in the beginning of a word and they are never found together in a single syllable. As mentioned earlier, the shape assumed by a character in a word is context sensitive; RNOON and RLAM both have all four shapes though they never come in the beginning of a word. RNOON and RLAM assume their initial shape when they come in the middle of a word and are followed by any character shown in figure 3. Figure 7 shows all four shapes of RNOON.

Figure 7: Initial, Middle, Final and independent shapes (from right to left) of RNOON

5.1.1. RNOON Problem

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

1. Plain NOON ن – let the sound be determined by context. This shape is the same as normal NOON and thus leads to ambiguity.
2. NOON with two vertical dots instead of one dot Æ (used by Punjabi Adabi Board). This shape takes too much vertical space and hard to write in words like KAF – KAF context and is also ambiguous with TEY (ت).
3. NOON with TOAY mark instead of dot ٹ (also used in Sindhi and available in Unicode). This shape is very similar to TTEY ٹ, especially when it comes at the start and middle of the word.
4. NOON with both dot and TOAY mark above ڻ. This shape takes too much vertical space and hard to write in words like KAF – KAF context.
5. NOON with "kundi" as in Pashto  . The variation used in this shape is not used for any other character.

6. NOON with small circle instead of dot Æ. Small circle is new mark not found in Punjabi and Urdu, which, though unambiguous, is unconventional. (Used by Saver International – monthly magazine and Punjabi Science Board)

Mainly four shapes 1, 2, 4 and 6 are being used in different books, periodicals and magazines. *Shape 1* is most widely used because of non availability of any publishing IT solution for Punjabi. *Shape 6* is totally unconventional. Only shape 2 and 4 have historical evidence to be used for RNOON (ڼ) as these two are found in different Punjabi manuscripts.

If we look at the history of development of Shahmukhi script, there is a tradition to build new characters from already existing characters by putting TOAY (ط) mark above it, e.g. TTEH (ٹ), DDAL (ڈ), RREH (ڑ) (these sounds are missing in Persian and Arabic scripts and are native to local languages of Subcontinent). Same tradition can also be applied to suggest a character for RNOON (ڻ).

If we look at DAL (د), ZAL (ذ) and DDAL (ڈ), these sounds are quite similar but harder from the previous one. A dot is placed to produce slight hard sound and TOAY (ط) mark to produce much harder sound. Same is the case with REH (ر), ZEH (ز) and RREH (ڑ). Similarly, in case of NOONGHUNA (ں), NOON (ن) and RNOON (ڼ), sound gets harder and harder from first to third character. So we can deduce by analogy that shape 4 should be used to represent RNOON. Also in Hindko, RNOON is written as a combination of NOON (ن) and RREH (ڑ). For example, پاݨی (water) is written as پانڑی. Thus TOAY (ط) mark of RREY (ڑ) is placed over NOON (ن) to suggest a character for RNOON (ڼ). Also this character has historical evidence to be used for RNOON through manuscripts. Acoustic analysis of RNOON [10] also shows that *shape 4* is the best representation of RNOON phoneme. So it is highly recommended that *shape 4* (ڼ) should be used for RNOON (ڼ).

5.1.2. RLAM Problem

In case of RLAM, there is a disagreement that it should be included in the character set or not. RLAM has its very rare usage in Punjabi, but definitely it is a sound of Punjabi. By definition, character set is a set of characters required to write a language completely [1], so RLAM should be included in the character set for Punjabi Shahmukhi. Also in Gurmukhi, RLAM was not assigned a character in the beginning, but later on character ਲ਼, derived by putting a dot below LALLAA (ਲ - character for LAM sound), was assigned to it to fulfill the language needs. As far as its shape is concerned, it is highly recommended that shape of LAM with TOAY mark above (ڵ) should be used for RLAM in the light of discussion for RNOON shape.

5.2. Aspirated Characters

HEH-DOACHASHMEE (ھ) is possibly not a character but a consonant modifier to indicate aspirated consonants. It combines with a variety of Punjabi consonants to give new consonants, e.g. ب + ھ = بھ and پ + ھ = پھ. There are sixteen aspirates in Punjabi. In figure 8, first line shows the aspirates in which aspirated character joins with HEH-DOACHASHMEE and second line shows the aspirates in which aspirated character does not join with HEH-DOACHASHMEE. Ten red boxed aspirates are very frequent in their usage, but others are very rare.

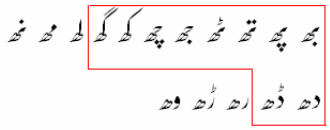


Figure 8: Punjabi Aspirates

The current system to represent the aspirates is working fairly well. Also in UZT 1.01 [11] (a standard for Urdu, approved by Government of Pakistan in July 2000), traditional way of writing aspirates is not changed. So it is highly recommended that only *Dochashmihay* (ھ) should be included into the character set to represent the aspirates.

5.3. Digits

All of Punjabi digits from zero through nine are already in Unicode from code point 06F0 to 06F9 except the Punjabi digits four, six and seven have different shapes as compared to given in Unicode. They are same as Urdu digits in UZT 1.01 [11]. I also have given images from Punjabi manuscripts [12], [13] to give the historical evidence of the usage of these digits in Punjabi. These digits are also being used in different Punjabi books [14], [15], [16]. It will be against tradition and we will not be able to reproduce manuscripts as well as other Punjabi book containing these digits precisely, if we change Punjabi digits with English digits. So it is highly recommended that Punjabi digits should not be substituted with Latin digits.

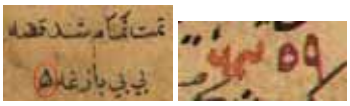


Figure 8: Images of Punjabi Digits from manuscript [12] and [13]

5.4. Diacritical Marks

Punjabi is very rich in diacritics and these diacritics are necessary to remove ambiguities in pronunciation and meaning of words. In simple words, whole vowel system of Punjabi depends on diacritics. All Punjabi diacritics are already present in Unicode except SUKUN (◌ْ). SUKUN is used to mark absence of a vowel after the base consonant. It has same effect as ARABIC SUKUN (already in Unicode at 0652 code) has. It also has been used in [12] and is shown in figure 9. So it is suggested that it should be included in the Punjabi Codepage as well as in Unicode. Most of Punjabi diacritics are used in manuscripts [12], [13], [18] and books [14], [15], [16].

Figure 9: SUKUN (circled) used in [12]

6. Proposed Character Set

After analysis of Punjabi characters, a character set with reference to ISO/IEC 10646/Unicode is proposed for Punjabi given in table 2 below. An exercise was done to identify Punjabi characters in Arabic block and corresponding Unicode are incorporated in the proposed codepage. As a result of this exercise, it was found that four characters do not have a representation in Unicode.

This code page is kept similar to ASCII code (where possible). This is because people are familiar with the character distribution in ASCII as it is a worldwide standard. In addition, owing to its universal acceptability, many hardware and software systems (especially the earlier ones, some of which are still deployed) conform very closely to ASCII standard. Incompatibility of this code page with ASCII would be mean incompatibility with these systems as well, which would not be a practical solution [11].

Table 2: Proposed Code Page of Punjabi – Shahmukhi

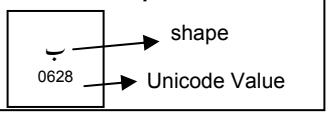
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			Sp 0020	۰ 06F0	@ 0040	ا 0627	ژ 0698	ں 06BA			ﷲ FDF2		[005B			
1			! 0021	۱ 06F1	HS ----	ب 0628	س 0633	ن 0646			◌ ----		\ 005C			
2			“ 0022	۲ 06F2	◌ ط 0615	پ 067E	ش 0634	ڻ ----			بسم الله الرحمن الرحيم FDFD		] 005D			
3			# 0023	۳ 06F3	◌ ً 064B	ت 062A	ص 0635	و 0648			ﷺ FDFA		US 005F			
4			Cr 0024	۴ 06F4	◌ ٌ 064C	ٹ 0679	ض 0636	ہ 06C1			◌ ص 0610		{ 007B			
5			٪ 066A	۵ 06F5	◌ ٍ 064D	ث 062B	ط 0637	ۃ 0629			◌ ع 0611		| 007C			
6			& 0026	۶ 06F6	◌ ّ 0651	ج 062C	ظ 0638	ء 0621			◌ ح 0612		} 007D			
7			‘ 0027	۷ 06F7	◌ ۓ 0553	چ 0686	ع 0639	ی 06CC			◌ ض 0613		Da 2013			
8			(0028	۸ 06F8	◌ ع 0654	ح 062D	غ 063A	ے 06D2			◌ 0614					
9			) 0029	۹ 06F9	◌ ع 0655	خ 062E	ف 0641	ھ 06BE			؏ 060F					
A			* 002A	: 003A	◌ ا 0656	د 062F	ق 0642	◌ َ 064E			ـ 0600					
B			+ 002B	؛ 061B	◌ ٰ 0657	ڈ 0688	ک 06A9	◌ ِ 0650			ـۂ 0601					
C			، 060C	< 003C	◌ ُ 0658	ذ 0630	گ 06AF	◌ ُ 064F			ـے 0602					
D			ـ 06D4	= 003D	◌ ٓ ----	ر 0631	ل 0644				ـہ 0603					
E			Dc ----	> 003E	◌ ا 0670	ڑ 0691	لٰ ----									→ ----
F			Dv 00F7 002F	؟ 061F	آ 0622	ز 0632	م 0645									

Abbreviations
Sp: Space, Cr: Currency, Dc: Decimal, Dv: Division, HS: Hard Space, US: Under Score, Da: Dash, →: Code plate switching

Legend

	Control Area (not to be used)
	Reserved Area (for future use)
	Vendor Area

Box Explaination



It is a 256 bit codepage and has been divided into various logical sections. It is divided into the following logical sections:

1. Control Characters (0 – 31, 127)
2. Punctuation and Arithmetic Symbols (32 – 47, 58 – 65)
3. Digits (48 – 57)
4. Punjabi Aerab/Diacritics (66 – 78, 123 – 125)
5. Punjabi Characters (79 – 121)
6. Reserved control space (128 – 159, 255)
7. Special Symbols (160 – 173, 192 – 199)
8. Reserved expansion Space (125,126, 174 – 191, 200 – 207, 240 – 253)
9. Vendor Area (208 – 239)
10. Toggle character (254)

Missing Punjabi characters have been listed in table 3. 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.

Table 3: Characters of Punjabi Codepage Proposed for Inclusion in ISO/IEC 10646/Unicode

Serial #	Symbol	Proposed Unicode	Proposed Description
1	ڵ	063B	ARABIC - PUNJABI SHAHMUKHI LETTER RLAM
2	ڼ	063C	ARABIC - PUNJABI SHAHMUKHI LETTER RNOON
3	◌ٜ	0659	ARABIC - PUNJABI SHAHMUKHI LETTER SUKUN
4	◌ؠ	0616	ARABIC LIGATURE JALLA JALALOUHOU

It should be noted that Unicode does not specify the collating sequence [17]. In case of Punjabi too, the collating sequence is defined through software. Unicode can serve as a source table for all the character and ligatures of Punjabi, as it does for other languages of the world.

7. Conclusion

ISO/IEC 10646/Unicode is fast assuming a standard for representing different national and local languages. After analysis of Punjabi character set, discussing and proposing solutions to different problems in the standardization of Punjabi character set, a codepage for Punjabi is proposed. Finally, comparison between this codepage and Unicode is done and a table of missing characters is drawn up. It is proposed that these characters be included in the Unicode standard.

Acknowledgement

I am very thankful to Dr. Khaver Zia for his full supervision in writing this paper. I am also very grateful to *Punjab Lok Sujag* who supported and helped me in doing this research.

Refference

[1] Muhammad Afzal and Sarmad Hussain (2001); “Urdu Computing Standards: Development of Urdu Zabata Takhti (UZT) 1.01 – WG2 N2413 – 2 – SC2 N3589 – 2”;

Proceedings of INMIC2001, Organized by IEEE & Lahore University of Management Sciences, Lahore, December 28 – 30, 2001, pp: 216 – 222.

[2] Tariq Rahman (2004); "Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift", Crossing the Digital Divide, SCALLA Conference on Computational Linguistics, 5th to 7th January 2004.

[3] Jones S et al (1992), A Digital Guide – Developing international user Information, International Edition, Digital Press, USA.

[4] Khaver Zia (1999) "Standard Code Table for Urdu", in the proceedings of 4th Symposium on Multilingual Information Processing (MLIT-4), Yangon, Myanmar, CICC, japan.

[5] 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 at FAST National University of Computer and Emerging Sciences, Lahore.

[6] http://www.wordiq.com/definition/Persian_language

[7] http://www.wordiq.com/definition/Gurmukhi

[8] Prof. Mirza Mqbul Baig Bdkhshani (1973); "Qavaid-e-Punjabi" Aziz Book Depu, Urdu Bazar, Lahore; Pg: 19 – 34 (in Punjabi).

[9] Shehbaz Malik (1996); "Punjabi Lisaniyat (Linguistics)" 3rd Ed. Aziz Book Depu, urdu Bazar, Lahore; Pg: 70 – 97 (in Punjabi).

[10] Tahira Naseem (2003), "Acoustic Analysis of Punjabi Phonemes /ɭ/ and /ɳ/" Akhbar-e-Urdu, June 2003, pp: 282 – 285.

[11] Muhammad Afzal and Sarmad Hussain (2001); "Urdu Computing Standards: Urdu Zabata Takhti (UZT) 1.01 – WG2 N2413 – 3 – SC2 N3589 – 3"; Proceedings of INMIC2001, Organized by IEEE & Lahore University of Management Sciences, Lahore, December 28 – 30, 2001, pp: 223 – 228.

[12] Hafiz Barkhurdar; "QISSA BIBI BAZGHA", Punjabi Manuscript in Oriental Section, Main Library University of the Punjab, Quaid-e-Azam Campus, Lahore Pakistan; 40 pages; Access # 8801.

[13] HASHIM SHAH (1899); "SOHNI MAHINWAL", Punjabi Manuscript in Oriental Section, Main Library University of the Punjab, Quaid-e-Azam Campus, Lahore Pakistan; 28 pages; Access # 8775.

[14] Muhammad Asif Khan (1992); "نِک سُک (Nik Suk)" (in Punjabi Shahmukhi); Pakistan Punjabi Adbi Board, Lahore.

[15] Damodar, arranged by Muhammad Asif Khan (1986); "ہیر دمودر (Heer Damodar)" (in Punjabi Shahmukhi); Pakistan Punjabi Adbi Board, Lahore.

[16] Baba Farid, arranged by Muhammad Asif Khan (1978); "آکھیا بابا فرید نے (Sayings of Baba Farid)" (in Punjabi Shahmukhi); Pakistan Punjabi Adbi Board, Lahore.

[17] Khaver Zia (1999); "Towards Unicode Standard for Urdu", in the Proceedings of 4th Symposium on Multilingual Information Processing (MLIT-4), Yangon, Myanmar, CICC, Japan.

[18] Maqbal (1747); "HEER RANJA", Punjabi Manuscript in Oriental Section, Main Library University of the Punjab, Quaid-e-Azam Campus, Lahore Pakistan; 55 pages; Access # 8791.