

# ANALYSIS OF NOORI NASTA'LEEQ FOR MAJOR PAKISTANI LANGUAGES

*M. G. Abbas Malik\**

*Christian Boitet\**

*Pushpak Bhattachariyya+*

Abbas.Malik@imag.fr

Christian.Boitet@imag.fr

pb@cse.iitb.ac.in

\*GETALP-LIG, Université de Grenoble (Ex. Université Joseph Fourier), France.  
+CSE, IIT Bombay, India.

## ABSTRACT

Nasta'leeq is a bidirectional, diagonal, non-monotonic, cursive, highly context-sensitive and very complex writing style for languages like Urdu, Punjabi, Balochi and Kashmiri. Each is written in a variant of the Perso-Arabic script. The style is characterized by well-formed orthographic rules that are passed down from generation to generation of calligraphers and old manuscripts. It is present in calligraphic arts and printed materials of the present, but orthographic rules have not been quantitatively analyzed in detail for the above-mentioned languages. This paper first presents the salient features of the Perso-Arabic script and briefly introduces its different writing styles. It also briefly discusses alphabets of major Pakistani languages. Finally, it gives the quantitative analysis of Nasta'leeq and explains its context-sensitive behavior with respect to Pakistani languages, knowing that it is equally true for Arabic, Persian and other languages written in derivations of the Perso-Arabic script. Finally, it discusses the Context-Sensitive Substitution Grammar of Nasta'leeq, a computational model of Nasta'leeq.

*Index Terms*— Nasta'leeq, script, Arabic, Persian, Urdu, Punjabi, Sindhi, Balochi, Kashmiri

## 1. INTRODUCTION

Pakistan is a country with at least six major languages and 58 minor ones [1]. Urdu, the national language, has over 11 million (7.57%) native speakers while those who use it as a second language are more than 105 million [2]. Punjabi, the mother tongue of 44.15% of the population, is the biggest language of Pakistan. Other major languages are Pashto, Sindhi, Balochi and Kashmiri. The size of these languages and Urdu is shown in Table 1.

The benefits from the Information Technology (IT) revolution cannot be reaped unless masses use it, which is not possible unless computing is possible in the languages that are understood by the masses [3]. Information has become such an integral part of our global society that access to it is considered as a basic human right. Internet is believed to be the dominant carrier of information across the globe. Currently, English is the lingua franca for Internet

and most of the information is available in it, but that makes information practically inaccessible to the vast majority of the world. This is applicable especially to countries like Pakistan where those who may be considered barely literate in Urdu represent only 43.92% population (66 millions according the 1998 census). That is rather a large number compared to the nearly 26 millions (17.29%) who, having passed the ten-year school system (matriculation), can presumably read and understand a little English. Internet and computer programs function in English in Pakistan and not even in Urdu let alone in the other languages. This means that most Pakistanis are either excluded from the digital world or function in it as handicapped aliens. In other words, Pakistani languages are under-resourced. Indeed, knowledge of English of most matriculates from Urdu and Sindhi medium schools is so rudimentary that they cannot carry out any meaningful interaction, especially those that would increase their knowledge or analytical skills, with the digital world. Perhaps only 4.38% graduates (about 6.5 millions) could do so [1].

Language	Number of Speakers
Urdu*	164,290,000
Punjabi	66,225,000
Pashto	23,130,000
Sindhi	21,150,000
Balochi	5,355,000
Kashmiri	4,496,000
* We include native and 2nd language speakers of Urdu. Source: [1]	

**Table 1: Speakers of Pakistani languages**

## 2. ARABIC SCRIPT AND ITS WRITING STYLES

The Arabic script is a cursive writing system. It has many writing styles, including Naskh, Kufi, Sulus, Riqah, Deevani, etc. Some of them are shown in figure 1. The Nasta'leeq writing style was developed in Iran during the 14<sup>th</sup> and 15<sup>th</sup> centuries by combining Naskh and Taleeq (an

old obsolete style)<sup>1</sup>. It is one of the main genres of the Islamic calligraphy. It is rich in calligraphic content. Owing to complexities of orthographic rendering, the basic shapes identified in this section are unable to render a language in an acceptable form in any Nasta'leeq style. A detailed quantitative analysis of Nasta'leeq with respect to Pakistani languages is given in section 4.

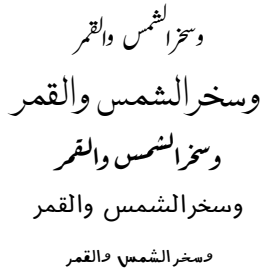


Figure 1: Different writing styles for Arabic

The distinguishing characteristics of Perso-Arabic script are discussed for the benefit of the unacquainted reader. It is read from right-to-left. Figure 2 shows some sample characters of Pakistani languages. Unlike English, characters do not have upper and lower case.

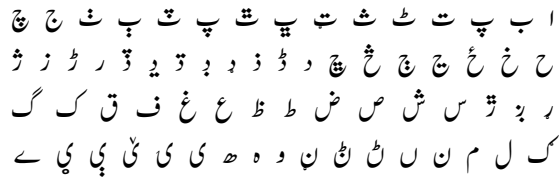


Figure 2: Sample characters of Pakistani languages

The shape assumed by a character in a word is context-sensitive, i.e. the shape is different depending on 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 [4,5]. Figure 3 shows these four shapes of the character Beh in Naskh writing style.

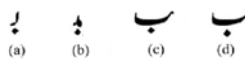


Figure 3: Context-sensitive shapes of BEH [4]

To be precise, the above is true for all except certain characters that only have the independent and the terminating shape when they occur at the beginning and the middle or end of a word respectively [4,5]. Some of these characters are shown in Figure 4.

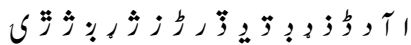


Figure 4: Sample characters having only two shapes

Hamza appears at the beginning of a word [4], but it could come at the beginning of a ligature. Also it takes the independent shape instead of the final shape when it comes

at the end of the word. Thus, it has initial, middle and independent shapes [4,5], as illustrated in figure 5.



Figure 5: Shapes of Hamza (circled) [5]

The Arabic, Persian and Pakistani languages have a large set of diacritical marks that are necessary for the correct articulation of a word. The diacritical marks appear above or below a character to define a vowel or to geminate a character [4,5]. They are the foundation of the vowel system in these scripts. The most common diacritical marks with the character Beh are shown in Figure 6.



Figure 6: BEH with Diacritical Marks

Diacritics, though part of the writing system, are sparingly used [4]. They are essential for ambiguities removal, natural language processing and speech synthesis [4,5,6,7].

### 3. PAKISTANI LANGUAGES

Pakistani languages are written in an alphabet that is derived from the Perso-Arabic alphabet. It is not possible to discuss all Pakistani languages here. This paper only discusses the six languages given in Table 1 because the last five represent the major geographical divisions of Pakistan, and Urdu is the National language of Pakistan. All of these languages belong to the Indo-European language family. Their family tree is given in Figure 7.

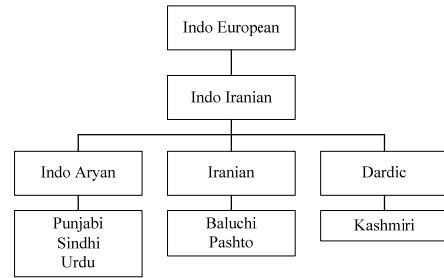


Figure 7: Language tree of 6 major Pakistani languages

The character sets of each of these languages are discussed separately here with their Unicode values. In Unicode, Arabic and its associated languages like Urdu, Punjabi, Pashto, Sindhi, etc. have been allocated the code points 0600h – 06FFh, 0750h – 077Fh and FB50h – FEFFh.

#### 3.1. Urdu

Urdu is the National language of Pakistan and one of the state languages of India with more than 60 million native speakers. It is one of the largest languages of the world, if one considers Hindi/Urdu as dialects of the same language called Hindustani by Platts [8]. Table 2 gives the size of Hindi/Urdu.

<sup>1</sup> [http://en.wikipedia.org/wiki/Nasta%27liq\\_script](http://en.wikipedia.org/wiki/Nasta%27liq_script)

Speakers	Native	2 <sup>nd</sup> Language	Total
Hindi	366,000,000	487,000,000	853,000,000
Urdu	60,290,000	104,000,000	164,290,000
Total	426,290,000	591,000,000	1,017,000,000

**Table 2: Hindi and Urdu speakers [7]**

Urdu is written in Nasta'leeq style. It has 35 consonant characters representing 27 consonant sounds as some consonant sounds are represented by two or more consonant characters, e.g. the sound 's' is represented by three different characters Seh (س), Seen (س) and Sad (س) [7]. Out of 35 consonant characters, 32 are adopted from Persian. 3 retroflex consonants are added to accommodate the indigenous sounds of the Indian sub-continent. These characters are Tteh (ٹ) [t̪], Ddal (ڈ) [d̪] and Rreh (ڑ) [ɽ]. Non-aspirated consonants of Urdu are given in Table 3.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ب [b]	0628	19	س [s]	0635
2	پ [p]	067E	20	ز [z]	0636
3	ت [t]	062A	21	ط [t̪]	0637
4	ث [t̪]	0679	22	ظ [z̪]	0638
5	ث [s]	06B2	23	ع [ʔ]	0639
6	ج [d͡ʒ]	062C	24	غ [ɣ]	063A
7	چ [t͡ʃ]	0686	25	ف [f]	0641
8	ح [h]	062D	26	ق [q]	0642
9	خ [x]	062E	27	ک [k]	06A9
10	د [d]	062F	28	گ [g]	06AF
11	ڈ [d̪]	0688	29	ل [l]	0644
12	ذ [z]	0630	30	م [m]	0645
13	ر [r]	0631	31	ن [n]	0646
14	ڑ [ɽ]	0691	32	و [v]	0648
15	ز [z]	0632	33	ه [h]	06C1
16	ژ [ʒ]	0698	34	ی [j]	06CC
17	س [s]	0633	35	ف [f]	0629
18	ش [ʃ]	0634			

**Table 3: Non-aspirated Urdu consonants**

The phenomenon of aspiration does not exist in Persian or Arabic but it exists in languages of the region e.g. Hindi, Urdu, Punjabi, etc. In Urdu, the special character Heh Doachashmee (ہ) is used to mark the aspiration. Thus aspirated consonants are represented by the combination of the consonant to be aspirated and Heh Doachashmee (ہ) e.g. ب [b] + ہ [h] = بھ [b<sup>h</sup>], ج [d͡ʒ] + ہ [h] = جھ [d͡ʒ<sup>h</sup>], etc. Urdu has 15 aspirated consonants [7]. Aspirated Urdu consonants are given in Table 4.

Sr.	Symbol	Sr.	Symbol	Sr.	Urdu
1	کھ [k <sup>h</sup> ]	6	گھ [t͡ʃ <sup>h</sup> ]	11	کھ [k <sup>h</sup> ]
2	پھ [p <sup>h</sup> ]	7	دھ [d̪ <sup>h</sup> ]	12	گھ [g <sup>h</sup> ]
3	تھ [t̪ <sup>h</sup> ]	8	ڈھ [d̪ <sup>h</sup> ]	13	لھ [l <sup>h</sup> ]
4	ٹھ [t̪ <sup>h</sup> ]	9	رھ [r <sup>h</sup> ]	14	مھ [m <sup>h</sup> ]
5	ثھ [s <sup>h</sup> ]	10	ڑھ [ɽ <sup>h</sup> ]	15	نھ [n <sup>h</sup> ]

**Table 4: Aspirated Urdu consonants**

In addition to consonants, Urdu has 10 vowels and 7 of them also have nasalized forms [9]. They are represented with the help of four long vowels (Alef Madda (ا), Alef (إ), Waw (و) and Yeh (ی)) and three short vowels (Arabic Fatha (َ), Damma (ُ) and Kasra (ِ)). The representation of a vowel is context-sensitive, i.e. a vowel may be written in two or more ways according to the context in a word, e.g. the vowel sound [ə] is represented by Alef (إ) + Zabar (ا) at the start of a word and by Zabar (ا) in the middle of a word. The vowel sound [ə] never comes at the end of a word. Nasalization of a vowel is marked with Noon-ghunna (ن) and with Noon (ن) at the end and in the middle of a word respectively [7]. For more details, see [7].

Urdu contains 15 diacritical marks. They represent vowel sounds, except Hamza-e-Izafat (ء) and Kasr-e-Izafat

(آ) that are used to build compound words, e.g. اورہ سائیس

[iɖrəhɪsɪms] (Institute of Science), تاریخ پیدائش [tarixipedaɪʃ] (date of birth), etc. Shadda (ّ) is used to geminate a

consonant e.g. رب [rəbb] (God), بھلا [əʃtʃʃʰa] (good), etc. Sukun (◌◌) is used to mark the absence of a vowel after the base consonant [7,8].

Pakistani languages also share the Perso-Arabic punctuation and special symbols. These punctuation marks and symbols are given in Table 5.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	،	060C	10	ع	060F
2	؛	061B	11	◌◌	0610
3	؟	061F	12	◌◌	0611
4	-	06D4	13	◌◌	0612
5	◌◌	0600	14	◌◌	0613
6	◌◌	0601	15	◌◌	0614
7	◌◌	0602	16	◌◌	0615
8	◌◌	0603	17	٪	066A
9	◌◌	060E			

**Table 5: Punctuation marks and other symbols**

Urdu has a numeral system that is derived from Persian. It assigns the same Unicode values as Persian ranging 06F0 – 06F9 but employs different shapes for number 4, 5 and 7. They are shown in Table 6.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	۰	06F0	6	۵	06F5
2	۱	06F1	7	۶	06F6
3	۲	06F2	8	۷	06F7
4	۳	06F3	9	۸	06F8
5	۴	06F4	10	۹	06F9

Table 6: Urdu numerals

### 3.2. Punjabi

Punjabi is written in two mutually incomprehensible scripts. One is the derivation of Perso-Arabic script (called Shahmukhi) used in Pakistan and the other is Gurmukhi, used in India. The Punjabi (Shahmukhi) alphabet is a superset of the Urdu alphabet and has one additional non-aspirated consonant, Rnoon (ڙ) [ɳ] [5,6]. The rest is the same as Urdu. Punjabi is also traditionally written in Nasta'leeq style. For more details on the Punjabi (Shahmukhi) alphabet see [5,6].

### 3.3. Pashto

Like Persian, Pashto does not have the aspiration. Heh Gol (ه) takes the shape of Heh Doachashmee (ه) when it comes at the start or middle of a ligature. Retroflex sounds also exist in Pashto like in Urdu and Punjabi, but Pashto employs different graphemes for them. Table 7 gives a shape comparison of retroflex consonants in six major Pakistani languages.

IPA	Urdu, Balochi, Kashmiri	Punjabi	Pashto	Sindhi
ʈ	ٹ	ਟ	ت	ت
ɖ	ڙ	ڑ	ڍ	ڍ
ʈʂ	ڙ	ڑ	ڍ	ڍ
ɳ	-	ڙ	ڙ	ڙ

Table 7: Comparison of retroflex consonants

In Pashto, there exist five different kinds of Yeh. One is employed as a consonant and the others represent different vowel sounds. They are shown in Figure 8.

ی [j], ی [i], ی [e], ی [əy], ی [ə]

Figure 8: Five Yehs of Pashto

Pashto has 39 consonants and uses the same Persian number system without any change. The vowel system of the Pashto script is also context-sensitive and is represented with the help of long vowels and diacritical marks. Pashto is traditionally written in Naskh style. Table 8 shows remaining Pashto characters that are not present in Urdu or have different shapes than in Urdu.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ځ [dz]	0681	4	ښ [ʒ]	069A
2	څ [ts]	0685	5	ګ [g]	06AB
3	ڙ [z]	0696			

Table 8: Pashto characters

### 3.4. Sindhi

Sindhi has 40 non-aspirated consonants and 11 aspirated consonants. In Sindhi, aspiration is expressed in different ways. For example, the aspiration of Jeem (ج) is indicated by Heh Doachashmee (ه) like in Urdu and Punjabi, and the aspiration of Beh (ب) is expressed by a separate new character with four dots below ڀ. Sindhi aspirated and non-aspirated consonants that are not present in Urdu or have different shapes from those in Urdu are given in Table 9.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ڀ [b]	067B	12	ڄ [d <sup>h</sup> ]	068D
2	ڀ [b <sup>h</sup> ]	0680	13	ڙ [ɳ]	0699
3	ڙ [t <sup>h</sup> ]	067F	14	ڙ [t <sup>h</sup> ]	-
4	ت [t]	067D	15	ڙ [p <sup>h</sup> ]	06A6
5	ت [t <sup>h</sup> ]	067A	16	ڪ [k]	06AA
6	ڙ [ ]	0684	17	ڪ [k <sup>h</sup> ]	06A9
7	ڙ [ɳ]	0683	18	ڙ [g]	06B3
8	ڙ [t <sup>h</sup> ]	0687	19	ڙ [ɳ]	06B1
9	ڙ [d <sup>h</sup> ]	068C	20	ڙ [ɳ]	06BB
10	ڙ [d]	068A	21	ڙ [j]	064A
11	ڙ [d]	068F			

Table 9: Aspirated and non-aspirated Sindhi consonants

Sindhi has 16 vowels that are also context-sensitive.

Pashto and Sindhi are both traditionally written in Naskh and their analysis for a Nasta'leeq style has never been done before. We have done it because they could also be written in Nasta'leeq just like Arabic<sup>2</sup>. Thus it is worthwhile to provide an analysis of Nasta'leeq for Pashto and Sindhi and provide an opportunity to the Pashto and Sindhi communities to write their languages in Nasta'leeq.

### 3.5. Balochi

Balochi uses a modified alphabet of Urdu and is written in Nasta'leeq style. Balochi has removed the redundant characters for the same sound, e.g. for the sound of [s], it keeps the character Seen (س) and discards the others (س, ش). Thus Balochi has 22 consonants. Like Persian and Pashto, it

<sup>2</sup> Arabic is also traditionally written in Naskh but there are very beautiful manuscripts of Arabic and Qur'an in the Indian sub-continent that are written in Nasta'leeq style. The first author has seen one in Pakistan.

also has no aspiration. It has two additional diacritics; one is the Hamza mark (◌ْ) above and the other is similar to the inverted Damma (◌ِ), but is horizontally reversed and much flatter (◌̣). Some native speakers also write Balochi using the Urdu script.

### 3.6. Kashmiri

Kashmiri employs the Urdu alphabet with a few additions to represent its specific vowels. Kashmiri has two additional Yehs (ي), one with an oval below (ي٘) and the other with a ‘v’ mark above (يٙ). It also has two additional Waws (و), one with a circle at the ending tail (و٘) and the other with a ‘v’ mark above (وٙ). In diacritical marks, it adds two diacritical marks, a slightly modified Hamza (ء) written above and below the character. The extra characters of Kashmiri are shown in Table 10. It is also traditionally written in Nasta’leeq style.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ي [ ]	-	4	ي٘ [e]	06CE
2	و [و]	06C4	5	وٙ [ə]	-
3	و٘ [o:]	06C6		وٙ [ ]	-

Table 10: Kashmiri characters

## 4. ANALYSIS OF NASTA’LEEQ

The rendering of Pakistani languages in Nasta’leeq is very complex because the shape of a character not only depends on its position (at the start, in the middle or at the end) in the word but also depends on surrounding characters in the word. The fundamental shapes of the analysis of Section 2 are not sufficient to produce orthographic rendering of major Pakistani languages in Nasta’leeq, because Nasta’leeq is inherently context-sensitive. Figure 9 shows different context-sensitive shapes of the character Beh.

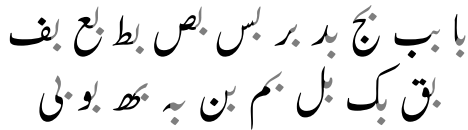


Figure 9: Context-sensitive shapes of Beh

Wali and Hussain [10] have given a quantitative analysis of Nasta’leeq (Nafees style) for Urdu. In this study, we give a quantitative analysis of the Noori style of Nasta’leeq for the six major Pakistani languages of Table 1.

For analysis purposes, we can divide our discussion in four parts, concerning independent shapes, two, three and four characters-joining. After the analysis of four characters long ligatures, the joining is recursive for ligatures longer than four, thus no further analysis and no new shapes are required to represent a text in Nasta’leeq style. This is shown in Figure 10.



Figure 10: Recursive nature of Nasta’leeq

To ease the analysis, we can divide characters into different groups on the basis of similarity in shapes. For example, the set of characters shown in Figure 11 can be grouped under the name Beh\_Family.



Figure 11: Beh\_Family members

The basic shape of each character of Figure 11 is exactly the same except their Noktas (dots or marks) above or below. Similarly, we can divide all other characters into different groups. All different groups of characters are given in Table 11.

Sr.	Name	Members
1	Alef	ا آ إ أ
2	Beh	ب پ ت ث ش ط
3	Jeem	ج چ خ ح
4	Dal	د ڈ ذ
5	Reh	ر ژ ز
6	Seen	س ش
7	Sad	ص ض
8	Toain	ظ
9	Ain	ع غ
10	Feh	ف ق
11	Qaf	ق
12	Kaf	ک گ
13	Lam	ل
14	Meem	م
15	Noon	ن ٹ
16	Waw	و و٘
17	Heh	ہ
18	Heh-Doachashmee	ھ
19	Hamza	ء
20	Choti-Yeh	ی ی٘ یٙ
21	Bari-Yeh	ے

Table 11: Character families

In addition to all characters of Table 11, there exist certain ligatures that are treated like independent characters in Nasta’leeq. They are given in Figure 12. They act like independent characters that do not join with the following character in the ligature and have only two (independent and final) shapes.

$$\begin{aligned} \text{گ} &= \text{ا} + \text{گ}, \text{ک} = \text{ا} + \text{ک}, \text{ل} = \text{ا} + \text{ل} \\ \text{گ} &= \text{ا} + \text{گ}, \text{ک} = \text{ا} + \text{ک}, \text{ل} = \text{ا} + \text{ل} \end{aligned}$$

Figure 12: Ligatures 1

#### 4.1. Independent Shapes

All characters of Table 11, the ligatures of Figure 12, the punctuation marks and the special symbols of Table 6, the Urdu Numerals of Table 5 and the Arabic numerals are independent characters. In addition to the punctuation marks of Table 6, other English punctuation marks like single quotes, double quotes, colon, etc. are also included in Nasta'leeq.

There are certain special ligatures that are included in Nasta'leeq, e.g. Allah ligature (الله), Muhammad ligature (محمد), etc. 23 other two character ligatures are also included in Nasta'leeq. In addition to all the above characters, Nasta'leeq also has a large set of diacritical marks that contains the diacritical marks of Arabic, Persian, Urdu, Punjabi, Pashto, Sindhi, Balochi, and Kashmiri. All these ligatures and diacritical marks are given in Table 12.

Sr.	Symbol	Sr.	Symbol	Sr.	Symbol
1	بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ	18	نے	35	ِ
2	الله	19	ٹے	36	ِ
3	محمد	20	پے	37	ِ
4	صَلَّىٰ عَلَیْهِمُ الصَّلَاةَ الْمُبَشِّرَةَ	21	پے	38	ِ
5	بے	22	ٹے	39	ِ
6	پے	23	پے	40	ِ
7	تے	24	ٹے	41	ِ
8	ٹے	25	ہ	42	ِ
9	ٹے	26	و	43	ِ
10	تے	27	ئی	44	ِ
11	پے	28	ِ	45	ِ
12	پے	29	ِ	46	ِ
13	تے	30	ِ	47	ِ
14	تے	31	ِ	48	ِ
15	نے	32	ِ	49	ِ
16	نے	33	ِ	50	ِ
17	ٹے	34	ِ	51	ِ

Table 12: Ligatures and diacritical marks

#### 4.2. Two Characters Joining

We do the analysis of two characters joining in reverse order, i.e. first we identify the final shape for an initial shape, a context before. The group having only two shapes consists of Alef, Dal, Reh, Waw, two characters from Choti-Yeh\_Family, ی (Alef Maskura) and ی (Pashto yeh with tail),

Bari-Yeh, La and Ka families. Some of these characters have two final shapes depending on their joining behavior with different families, e.g. Reh\_Family has two final shapes, one shape has only two (independent and final) shapes for Beh, Jeem, Kaf, Lam, Noon, Hamza and choti-yeh families and the other for the rest. The final shapes of 2-shapes families are given in Table 13.

Sr.	Shape	Examples
1	ا	پا چا سا صا نایا
2	ر	پد چد سد صد ند ید
3	ر ر	پر چر سر صر نر یر
4	و و	پو چو سو صو نو یو
5	ن ن	پنی چنی سی صنی نی ینی
6	ے	پے چے سے صے ٹے
7	لا	پلا چلا سلا صلا نلا یلا
8	گا کا	پگا چگا سگا صگا نا یگا

Table 13: Final shapes of Alef, Dal, Reh, Waw, Bari-yeh, La, Ka and two Choti-Yehs

Final shapes of the other families are given in Table 14.

Sr.	Shape	Examples
1	ب	بٹ بٹ بٹ بٹ بٹ بٹ
2	ج	چچ چچ چچ چچ چچ
3	س	بس جس سس صس نس یس
4	ص	بص حص صص نص یص
5	ط	بٹ جٹ سٹ صٹ ٹٹ یٹ
6	ع	بع جع سع صع نع یع
7	ف	بف جف صف صف نف یف
8	ق	بق جق صق صق نق یق
9	ک	بک چک سک صک نک یک
10	ل	بل جل سل صل نل یل
11	م	بم جم صم صم نم یم
12	ن	بن جن صن نن ین
13	-	بہ جہ صہ صہ نہ یہ
14	ھ	بھ جھ صھ صھ نہ ھ

Table 14: Final shapes

Hamza (ء) does not have a final shape. Thus there are 22 final families depending upon their final shapes, given in Table 13 and 14.

The above two tables not only give us the final shapes of all the families of Table 11 and of the ligatures of Figure 12 (La ۱ and Ka family گ, ک, کا, گ), they also give us the analysis of initial shapes of the Beh, Jeem, Seen, Sad, Noon and Choti-yeh families. The analysis of initial shapes of

Beh, Noon, Hamza and Choti-yeh family in the above examples shows that they have the same base form for the initial shape with variations in Noktas. It is also clear that the initial form for final shapes of the Sad and Ain families are the same. Thus the Behinit family (including initial forms of Beh, Noon, Hamza and Choti-yeh families) has 21 initial shapes. The initial shapes of the Behinit and Jeeminit families are given in Table 15.

Sr.	Behinit Shape	Jeeminit Shapes	Final Families
1	◌	◌	Alef_Final
2	◌	◌	Beh_Final
3	◌	◌	Jeem_Final
4	◌	◌	Dal_Final
5	◌	◌	Reh_Final
6	◌	◌	Seen_Final
7	◌	◌	Sad_Ain_Final
8	◌	◌	Tah_Final
9	◌	◌	Feh_Final
10	◌	◌	Qaf_Final
11	◌	◌	Kaf_Final
12	◌	◌	Lam_Final
13	◌	◌	Meem_Final
14	◌	◌	Noon_Final
15	◌	◌	Waw_Final
16	◌	◌	Hehgol_Final
17	◌	◌	Heh-doachashmee_Final
18	◌	◌	Choti-Yeh_Final
19	*	◌	Bari-yeh_Final
20	◌	◌	La_Final
21	◌	◌	Ka_Final

\* Behinit family with Bari-yeh is stored as ligatures

**Table 15: Initial shapes of Beh and Jeem families**

With 21 initial shapes of all families, all possible two character ligatures can be represented in Nasta'leeq. The Kaf and Lam families do not have an initial shape for Alef because these pairs are stored as ligatures, as shown in Figure 12.





### 4.3. Three Characters Joining

The final shapes have already been identified in the previous section. Similar to the initial shapes, 21 medial shapes are identified for the final shape families. The medial shapes of Behmedi and Jeemmedi families for final families are given in Table 16.

Sr.	Behmedi Shape	Jeemedi Shapes	Final Families
1	◌	◌	Alef_Final
2	◌	◌	Beh_Final
3	◌	◌	Jeem_Final
4	◌	◌	Dal_Final
5	◌	◌	Reh_Final
6	◌	◌	Seen_Final
7	◌	◌	Sad_Ain_Final
8	◌	◌	Tah_Final
9	◌	◌	Feh_Final
10	◌	◌	Qaf_Final
11	◌	◌	Kaf_Final, Gaf_Final
12	◌	◌	Lam_Final
13	◌	◌	Meem_Final
14	◌	◌	Noon_Final
15	◌	◌	Waw_Final
16	◌	◌	Hehgol_Final
17	◌	◌	Heh-doachashmee_Final
18	◌	◌	Choti-Yeh_Final
19	◌	◌	Bari-yeh_Final
20	◌	◌	La_Final
21	◌	◌	Ka_Final

**Table 16: Medial shapes of Beh and Jeem families**

The Behmedi shapes can be grouped into four different families according to their joining behavior with the previous character. This is shown in Table 17.

Name of Family	Shape	Members
Behmedi1		1, 2, 4, 7, 8, 9, 10, 11, 12, 15, 16, 19, 20, 21, 24, 25, 28, 29
Behmedi2		3, 13, 17, 18, 26, 30
Behmedi3		6, 14, 22, 23, 27
Behmedi4		5

**Table 17: Behmedi families**

For the families of Table 17, we need four initial shapes of each family having an initial shape. Thus the Behinit family has four new shapes for the Behmedi family, one shape for the Jeemmedi family and so on. All additional initial shapes of the Behinit and Jeeminit families, identified for medial shapes, are given in Table 18.

Sr.	Behinit Shape	Jeeminit Shapes	Medial Families
22	·	ˆ	Behmedi1
23	·	ˆ	Behmedi2
24	ˆ	ˆ	Behmedi3
25	·	ˆ	Behmedi4
26	ˆ	ˆ	Jeemmedi
27	ˆ	ˆ	Seenmedi
28	ˆ	ˆ	Sadmedi, Tahmedi, Ainmedi, Fehmedi
29	ˆ	ˆ	Kafmedi, Gafmedi, Lammedi
30	ˆ	ˆ	Meemmedi, Hehbolmedi, Heh-doachashmeemedi

**Table 18: More initial shapes of Beh and Jeem families**

We have thus 30 initial shapes and 21 medial shapes that represent all possible ligatures of length three of the six major Pakistani languages when written in the Noori Nasta'leeq style. It is not possible to list all shapes of all characters due to space shortage.

#### 4.4. Four Characters Joining

We do our analysis in the reverse direction, i.e. from left-to-right. In the analysis of three characters joining, we have already identified the shapes of the last two characters of our ligatures of length 4 that are final shapes and medial shapes for our final shapes. Now first we need to identify the medial shapes that will join with the already identified medial shapes. Secondly, we need to identify the initial shapes that will join with newly identified medial shapes in the previous step and this will complete our joining analysis.

Sr.	Behinit Shape	Jeeminit Shapes	Medial Families
22	·	ˆ	Behmedi1
23	·	ˆ	Behmedi2
24	ˆ	ˆ	Behmedi3
25	·	ˆ	Behmedi4
26	ˆ	ˆ	Jeemmedi
27	·	ˆ	Seenmedi
28	ˆ	ˆ	Sadmedi, Tahmedi, Ainmedi, Fehmedi
29	ˆ	ˆ	Kafmedi, Gafmedi, Lammedi
30	ˆ	ˆ	Meemmedi, Hehbolmedi, Heh-doachashmeemedi

**Table 19: More medial shapes of Beh and Jeem families**

The process of identifying the new medial shapes is the same as that used to identify the initial shapes for the first 21 medial shapes. Similar to the Behinit family, the Behmedi family also has four new shapes for its first 21 members, one shape for the Jeemmedi family and so on. All additional medial shapes of the Behmedi and Jeemmedi families, identified for medial shapes, are given in Table 19.

Table 17 shows that the Behmedi2 family includes the medial shapes # 26 and 30. Thus, fortunately, we do not have new initial shapes for these newly identified medial shapes of Table 19. Hence, our analysis for Noori Nasta'leeq style is complete.

Ligatures longer than 4 can be built using recursively the shapes already identified. That is shown in Figure 10. We have 1 or 2 final shapes, 30 initial shapes and 30 medial shapes for the characters of major Pakistani languages. Thus we need more than 1300 glyphs to represent the scripts of major Pakistani languages in the Noori Nasta'leeq style or build a good looking font for these languages.

## 5. CONTEXT-SENSITIVE SUBSTITUTION GRAMMAR

The analysis given in Section 4 can be represented in the *Context-Sensitive Substitution Grammar*. Figure 13 shows some rules of the contextual substitution grammar of Nasta'leeq.

### Initial Rule

beh → behinit1 aiknoktabelow  
jeem → jeeminit1 aiknoktabelow  
*No Context* (Before | After)

### Medial Rule

Beh → behmedi1 aiknoktabelow  
Jeem → jeemmedi1 aiknoktabelow  
*No Context* (Before | After)

### Final Rule

beh → behfinal  
jeem → jeemfinal  
*No Context* (Before | After)

### Contextual Substitution Rule for Behfinal

behinit1 → behinit2  
jeeminit1 → jeeminit2  
behmedi1 → behmedi2  
jeemmedi1 → jeemmedi2  
*Context* (| behfinal)

### Contextual Substitution Rule for Jeemfinal

behinit1 → behinit3  
jeeminit1 → jeeminit3  
behmedi1 → behmedi3  
jeemmedi1 → jeemmedi3  
*Context* (| jeemfinal)

### Contextual Substitution Rule for Behmedi1 Family

behinit1 → behinit22  
jeeminit1 → jeeminit22  
behmedi1 → behmedi22  
jeemmedi1 → jeemmedi22  
*Context* (| <behmedi1 Family>)

**Figure 13: Context-Sensitive Substitution Grammar**



The Initial Rule tells that Beh (ب) and Jeem (ج) are substituted by behinit1 (.) and jeeminit1 (.) respectively with appropriate Nokta on them whenever they come at the initial position of a ligature. Medial and Final rules also have the same kind of interpretation for the medial and final positions respectively. The Contextual Substitution Rule for Behfinal1 tells that default initial shapes behinit1 (.) and jeeminit1 (.) at the initial position are substituted by behinit2 (-) and jeeminit2 (.) when they are followed by a glyph of the Behfinal1 family. It also tells that default medial shapes behmedi1 (.) and jeemmedi1 (.) at the medial position are substituted with behmedi2 (-) and jeeminit2 (.) when they are followed by a character of the Behfinal1 family. The other rules have the same kind of interpretations. Figure 13 shows a very small part of the *Context-Sensitive Substitution Grammar* of Noori Nasta'leeq. This shows the contextual nature and complexity of the Noori Nasta'leeq style. Theoretically, the *Context-Sensitive Substitution Grammar* is a computational model of the Noori Nasta'leeq contextual complexity.

## 6. CONCLUSION

Nasta'leeq is a bidirectional, diagonal, non-monotonic, cursive, highly context-sensitive and very complex writing system for languages written in the Arabic or in extended Arabic scripts like those of Urdu, Punjabi, Pashto, Sindhi, Balochi, Kashmiri, etc. The analysis of Nasta'leeq for major Pakistani languages applies equally to Arabic, Persian and other languages written in extended Arabic scripts. The analysis of Nasta'leeq and the *Context-Sensitive Substitution Grammar*, discussed in this paper, can be used to build a good quality and high speed font for the Arabic, Persian, Urdu, Punjabi, Pashto, Sindhi, Balochi and Kashmiri languages to write them in the Noori Nasta'leeq style.

The practical implementation of a character-based Nasta'leeq font for Arabic, Persian and Pakistani languages is a much more complex process than its theoretical analysis. A practical development of a Nasta'leeq font not only needs the *Context-Sensitive Substitution Grammar*, but it also requires other important and vital positioning information to correctly position glyphs and Noktas considering their contextual glyphs and Noktas, as shown in Figure 10. An implementation for Urdu and Punjabi has been produced by the first author in 2004 and is available as a freeware on the Web at [www.puran.info](http://www.puran.info)<sup>3</sup>. Practical details cannot be discussed here due to shortage of space. We plan to discuss it in a future paper. Digital graphical representation in a computer is vital for under-resourced languages, so that native people can understand their native languages and can contribute to the development of computational linguistic resources for their languages.

## 7. REFERENCES

- [1] Rahman, T. “*Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift*”, in proc. Crossing the Digital Divide, SCALLA Conference on Computational Linguistics, pp. 5-7 January, 2004.
- [2] Grimes, B. F. “*Pakistan*”. Ethnologue: Languages of the World. 14th Edition Dallas, Texas; Summer Institute of Linguistics, 2000.
- [3] Afzal, M. and Hussain, S. “*Urdu Computing Standards: Development of Urdu Zabta Takhti (UZT) 1.01*”. in proc. INMIC-2001, Lahore, 2001.
- [4] Khaver, Z. “*Standard Code Table for Urdu*”, in proc. 4th Symposium on Multilingual Information Processing (MLIT-4), Yangon, Myanmar, CICC, Japan, 1999.
- [5] Malik, M. G. Abbas; “*Towards a Unicode Compatible Punjabi Character Set*”. In proc. 27<sup>th</sup> Internationalization and Unicode Conference, Berlin, Germany, 2005.
- [6] Malik, M. G. Abbas; “*Punjabi Machine Transliteration*”. In proc. 21<sup>st</sup> International Conference on Computational Linguistics COLING-06 and 44<sup>th</sup> Annual Meeting of ACL, Sydney, Australia, 2006.
- [7] Malik, M. G. Abbas; Boitet, Christian; and Bhattacharyya, Pushpak; “*Hindi Urdu Machine Transliteration using Finite-state Transducers*”. In proc. 22<sup>nd</sup> International Conference on Computational Linguistics COLING-08, Manchester, UK, 2008.
- [8] Platts, J. T. A *Grammar of the Hindustani or Urdu Language*. Crosby Lockwood and Son, 7 Stationers Hall Court, Ludgate hill, London. E.C., 1909.
- [9] Hussain, S. “*Letter to Sound Rules for Urdu Text to Speech System*”, in Proc. of Workshop on “Computational Approaches to Arabic Script-based Languages”, COLING-04, Geneva, Switzerland, 2004.
- [10] Wali, A., Hussain, S., “*Context Sensitive Shape-Substitution in Nastaliq Writing System: an analysis and fomulation*”. In Proc. of International Joint Conferences on Computer, Information and Systems Sciences and Engineering, 2006.

<sup>3</sup> [www.puran.info](http://www.puran.info)