Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation

Aamir Wali
University of Illinois at Urbana-Champaign
awali2@uiuc.edu

Sarmad Hussain
NUCES, Lahore
sarmad.hussain@nu.edu.pk

***Abstract*- Urdu is a widely used language in South Asia and is spoken in more than 20 countries. In writing, Urdu is traditionally written in Nastaliq script. Though this script is defined by well-formed rules, passed down mainly through generations of calligraphers, than books etc, these rules have not been quantitatively examined and published in enough detail. The extreme context sensitive nature of Nastaliq is generally accepted by its writers without the need to actually explore this hypothesis. This paper aims to show both. It first performs a quantitative analysis of Nastaliq and then explains its contextual behavior. This behavior is captured in the form of a context sensitive grammar. This computational model could serve as a first step towards electronic Typography of Nastaliq.**

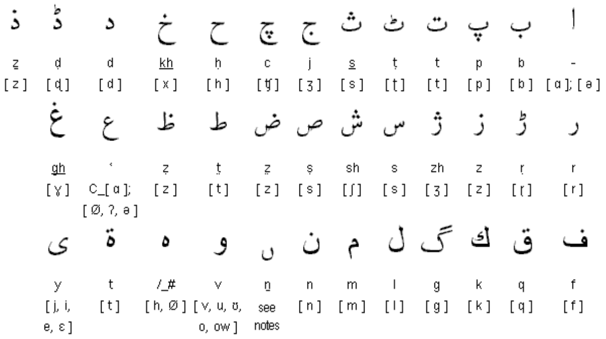
I. Introduction

Urdu is spoken by more than 60 million speakers in over 20 countries [1]. Urdu is derived from Arabic script. Arabic has many writing styles including Naskh, Sulus, Riqah and Deevani. Urdu however is written in Nastaliq script which is a mixture of Naskh and an old obsolete Taleeq styles. This is far more complex than the others.

Firstly, letters are written using a flat nib (traditionally using bamboo pens) and both trajectory of the pen and angle of the nib define a glyph representing a letter. Each letter has precise writing rules, relative to the length of the flat nib. Secondly, this cursive font is highly context sensitive. Shape of a letter depends on multiple neighboring characters. In addition it has a complex mark placement and justification mechanism. This paper examines the context sensitive behavior of this script and presents a context sensitive grammar explaining it.

A. Urdu Script

The Urdu abjad is a derivative of the Persian alphabet derived from Arabic script, which in itself is derived from the Aramaic script (Encarta 2000, Encyclopedia of Writing and [2]). Urdu has also retained its Persio-Arabic influence in the form of the writing style or typeface. Urdu is written in Nastaliq, a commonly used calligraphic style for Persio-Arabic scripts. Nastaliq is derived from two other styles of Arabic script 'Naskh' and 'Taleeq'. It was therefore named Naskh-Taleeq which gradually shortened to "Nastaliq".



ذ	ڈ	د	خ	ح	چ	ج	ث	ٹ	ت	پ	ب	ا
z̤	ḍ	d	kh	ḥ	c	j	s̱	ṭ	t	p	b	-
[z]	[ɖ]	[d]	[x]	[h]	[ʧ]	[ʒ]	[s]	[ʈ]	[t]	[p]	[b]	[ɑ]; [ə]
	غ	ع	ظ	ط	ض	ص	ش	س	ژ	ز	ڑ	ر
	gh	ʿ	ẓ	ṭ	ẕ	ṣ	sh	s	zh	z	ṛ	r
	[ɣ]	C_[ɑ]; [Ø, ʔ, ə]	[z]	[t]	[z]	[s]	[ʃ]	[s]	[ʒ]	[z]	[ɽ]	[r]
	ی	ة	ہ	و	ں	ن	م	ل	گ	ك	ق	ف
	y	t	/_#	v	n̠	n	m	l	g	k	q	f
	[j, i, e, ɛ]	[t]	[h, Ø]	[v, u, ʊ, o, ow]	see notes	[n]	[m]	[l]	[g]	[k]	[q]	[f]

Fig. 1. Urdu Abjad

II. Positional And Contextual Forms

Arabic is a cursive script in which successive letters join together. A letter can therefore have four forms depending on its location or position in a ligature. These are isolated, initial, medial and final forms. Consider the following table 1, in which letter 'bay' indicated in gray has a different shape when it occurs in a) initial, b) medial, c) final and d) isolated position. Since Urdu is an derived from Arabic script and Nastaliq is used for writing Urdu, both Urdu and Nastaliq inherit this property.

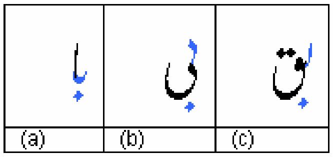
TABLE 1
Positional Forms For Letter *Bay*

با	قبا	قب	ب
(a)	(b)	(c)	(d)

Letters 'alif', 'dal', 'ray' and 'vao' only have two forms. These letters cannot join from front with the next letter and therefore do not have an initial or medial forms.

Nastaliq is far more complex than the 4-shape phenomenon. In addition to position of character in a ligature, the character shape also depends on other characters of the ligature. Thus Nastaliq is inherently context sensitive. Table 2 below shows a sample of this behavior in which a letter bay, occurring in initial form in all cases, has three different shape indicated in grey. This context sensitivity of Nastaliq can be captured by substitution grammar. This is discussed in detail later in this paper.

TABLE 2
CONTEXTUAL FORMS FOR LETTER *BAY* IN INITIAL FORM



III. GROUPING OF 'SIMILAR' LETTERS

There are some letters in Arabic script and consequently in Urdu, that share a common base form. What they differ by is a diacritical mark placed below or above the base form. This can be seen in table 3 below which shows letters 'bay', 'pay', 'tay', 'Tey' and 'say' in isolated forms. And it's clearly evident that they all have the same base form. This is also true for initial, medial and final forms of these letters.

TABLE 3
LETTERS WITH SIMILAR BASE FORM

Isolated Form	ب	پ	ت	ٹ	ث
Initial form	بـ	پـ	تـ	ٹـ	ثـ
	(a)	(b)	(c)	(d)	(e)

Since these letters have a similar base shape, it would be redundant to examine the shape of all these letters in different positions and context. Studying the behavior of one letter would suffice the others. Table 4 below shows all groupings that are possible. The benefit of this grouping is that instead of examining about 35 letters in Urdu, only half of them need to be looked into. Note that only the characters that are used in place of multiple similar shapes are shown. The rest of the characters in the abjad are used without any such similar-shape classification.

TABLE 4
GROUPING OF LETTERS WITH SIMILAR BASE FORM

Similar Base Forms	Letter
ث ٹ ت پ ب Also ں and ی in initial and medial form	ب
خ ح چ ج	ج
ذ ڈ د	د
ژ ز ڑ ر	ر
ش س	س
ض ص	ص
ظ ط	ط
غ ع	ع
گ ک	ک

IV. METHODOLOGY: TABLETS

The Nastaliq alphabets for Urdu have been adapted from their Arabic counterparts as in the Naskh and T'aleeq styles from which it has been derived. However, even for Urdu, this style is still taught with its original alphabet set. When the pupil gains mastery of the ligatures of this alphabet, then he/she is introduced to the modifications for Urdu.

The methodology employed for this study is similar to how calligraphy is taught to freshmen. The students begin by writing isolated forms of letters. In doing so, they must develop the skill to write a perfect shape over and over again by maintaining the exact size, angle, position etc. When the students have achieved the proficiency in isolated form it is said that they have completed the first 'taxti', meaning tablet. The first tablet is shown in figure 2 below; 'taxti' or tablet can be considered as a degree of excellence. First tablet is considered level 0

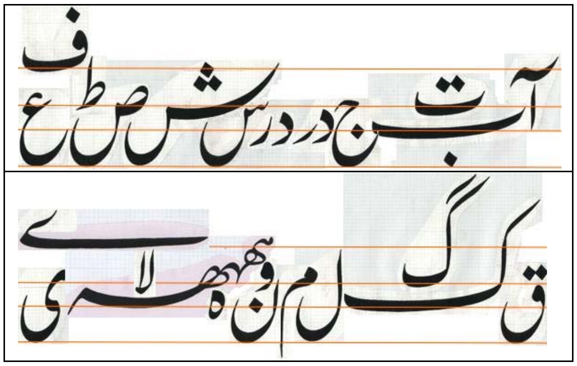


Fig. 2. Tablet for Isolated forms

Once this level is completed, the students move over to level 2. In this level, they learn to write all possible two-letter ligatures. This phase is organized into 10 tablets. The first tablet consists of all ligatures beginning with 'bay', the second abjad of Arabic script. For example in English it would be like writing ba, bb, bc etc. all the way till bz. Note that the first letter in Arabic script is 'alif' which has no initial form and therefore does not form two letter ligatures beginning with it. See section 2 above. The bay tablet is shown in figure 3 below.

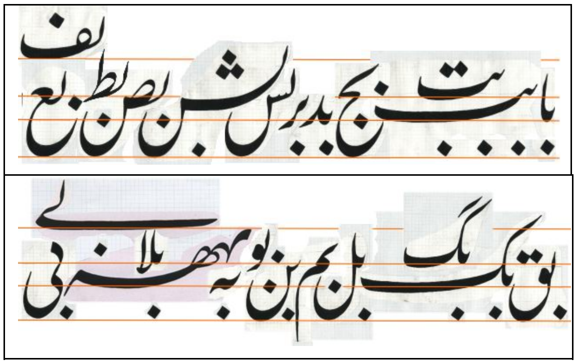


Fig. 3. Bay Tablet

In the similar way the students move over to the second tablet of this phase which has ligatures beginning with 'jeem' and so on. The 'jeem' tablet is also shown below in figure 4.

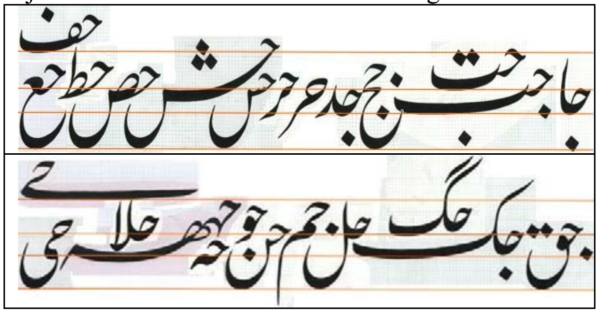


Fig. 4. Jeem Tablet

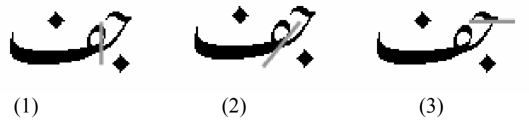
Note: If the students learn to write ligatures beginning with 'bay', then it is automatically assumed they can also write ligatures beginning with 'tay', 'say', etc because all these letters have the same shape as 'bay' and only differ in the diacritical mark. Please refer to section 3 above for details.

The level 0 and 1 helps to understand a lot about the initial and final forms of letters. There are however no further (formal) levels of any kind. So, in addition to understanding the medial forms, the students are also expected to learn to write three and above letter ligatures on their own through observation and consultation.

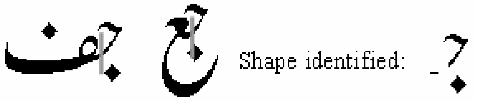
For this study, a calligrapher was consulted who provided the tablets and other text that was required.

V. Segmentation

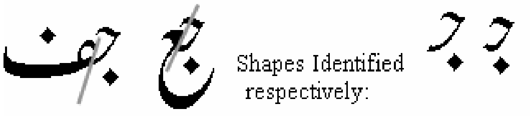
Before moving over to the analysis of Contextual substitution of Nastaliq, one of the predecessor worth mentioning is segmentation. Since Nastaliq is a cursive script, segmentation plays an important role in determining the distinct shapes a letter can acquire. Consider a two letter ligature composed of jeem and fay. There can be a number of places from where this ligature can be segmented (as indicated by 1,2 and 3) giving different shapes for initial form of jeem.



If (1) is adopted as the segmentation scheme and being consistent with it, when this is applied on another two letter ligature that also begins with a 'jeem', it follows that both the ligatures share a common shape of 'jeem'. Both the ligature are shown below with segmentation approach (1). The resulting, similar 'jeem' shapes is also given.



Where as if the second (2) option is accepted for segmentation, then the two initial forms will be different.



The segmentation approach selected for this analysis is the later one or approach (2). The reason for this selection is that the resulting shapes are a close approximation to the shapes in the calligrapher's mind. Most probably because these shapes represent complete stokes (, though an expert calligrapher may be able to write the whole ligature in one stroke). While in former case there seems to be some discontinuity in the smooth stroke of a calligrapher's pen.

The discussion on contextual shape analysis in next section is therefore based on the "stroke segmentation' approach.

VI. Contextual Analysis Of Nastaliq

This section lists the context sensitive grammar for characters occurring in initial position of ligatures.

Explanation of Grammatical Conventions:

The productions such as:

ب → ب$_1$ / __<A> | ف $_{\text{Initial Form}}$ __

is to be read as ب transforms to ب$_1$ (ب → ب$_1$), in the environment (/) when ب occurs before class A (__<A>) or (|) when ب occurs before initially occurring ف.

Note that 'OR' (|)operator has a higher precedence than 'Forward Slash' (/) operator. Thus, it would be possible to write multiple transformations in one environment using several 'OR' (|) operator on the right side of a single (/).

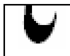
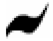
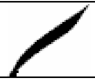
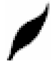
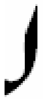
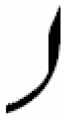
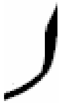
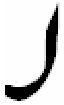
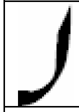
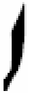
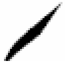
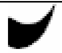
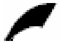
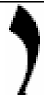
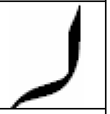
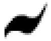
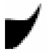
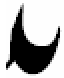
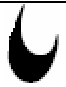
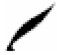
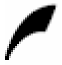
Contextual Shift

One invariant that have predominantly existed in this contextual analysis of Nastaliq is that *the shape of a character is mostly dependent on immediate proceding character.* That is given a ligature composed of character sequence $X_{1,}$ X_2 ….X_N for N>2, the shape of character X_i where i < N, is determined by letter $X_{i+1.}$ While all preceding letters X_1 ....X_{i-1} and character sequences after its following character i.e X_{i+2} …. X_N have no (or little) role in its shaping. Sequence of bay's form an exception to this general rule. Other exceptions are also mentioned

Initial-Position characters:

The following table lists the initial shapes of letter 'Bay'. The last eight shapes were identified during analysis of three-character ligature and do not realize in two character ligatures. Another way of putting it is that these form can only occur in ligatures of length greater than or equal to three. There is no such restriction for the first 16 shapes.

TABLE 5
INITIAL FORMS OF BAY

			
ب$_{Init1}$	ب$_{Init2}$	ب$_{Init3}$	ب$_{Init4}$
			
ب$_{Init5}$	ب$_{Init6}$	ب$_{Init7}$	ب$_{Init8}$
			
ب$_{Init9}$	ب$_{Init10}$	ب$_{Init11}$	ب$_{Init12}$
			
ب$_{Init13}$	ب$_{Init14}$	ب$_{Init15}$	ب$_{Init16}$
			
ب$_{Init21}$	ب $_{Init22}$	ب $_{Init23}$	ب$_{Init24}$
			
ب$_{Init25}$	ب$_{Init26}$	ب$_{Init27}$	ب$_{Init28}$

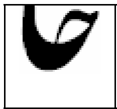
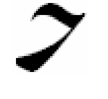
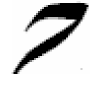
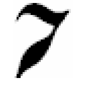
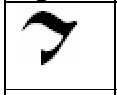
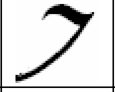
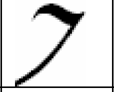
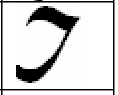
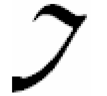
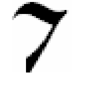
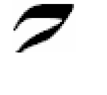
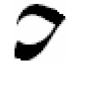
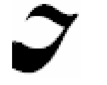
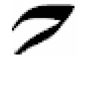
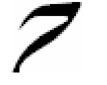
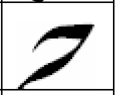
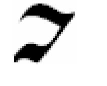
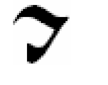
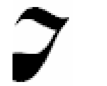
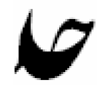
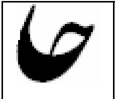
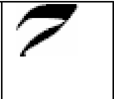
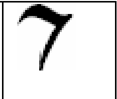
The context in which these shapes occur has been formulated in the form of a context sensitive grammar. The grammar for initial forms of 'bay' is given below. Note: The word 'medial' used in the grammar represents medial shapes of 'bay' given in appendix A

ب → ب $_{Init1}$ / __ ل ___ / ک ___ | د ___ | ا
ب → ب $_{Init2}$ / __ ب$_{Final}$
ب → ب $_{Init3}$ / __ ج
ب → ب $_{Init4}$ / __ ر
ب → ب $_{Init5}$ / __ س
ب → ب $_{Init6}$ / __ ص
ب → ب $_{Init7}$ / __ ط

ب → ب Init8 / ___ ع
ب → ب Init9 / ___ ف
ب → ب Init10 / ___ ق | ___ و
ب → ب Init11 / ___ م
ب → ب Init12 / ___ ن
ب → ب Init13 / ___ ہFinal
ب → ب Init14 / ___ ھ
ب → ب Init15 / ___ ی
ب → ب Init16 / ___ ے
ب → ب Init22 /___ب Medial Only if none of the below apply
ب → ب Init21 /___ب Medial2
ب → ب Init23 /___ب Medial4
ب → ب Init24 /___ب Medial5
ب → ب Init25 /___ب Medial13
ب → ب Init26 /د م ___ | ا م ___
/ ___ل م ___ / ک م
ب → ب Init27 /___م Otherwise
ب → ب Init28 /___ہ Medial

All other letters have similar number of shapes occurring in similar context. The initial forms of 'jeem' have been shown in table below. The grammar of 'jeem' can be derived from grammar of 'bay'. Likewise, shapes of other letters can be deduced from table 5 and table 6. Medial Shapes of Bay have been listed in Appendix A

TABLE 6
INITIAL FORMS OF JEEM

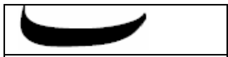
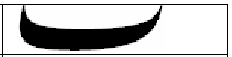
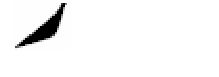
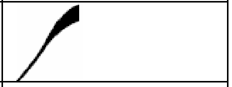
			
ج Init1	ج Init2	ج Init3	ج Init4
			
ج Init5	ج Init6	ج Init7	ج Init8
			
ج Init9	ج Init10	ج Init11	ج Init12
			
ج Init13	ج Init14	ج Init15	ج Init16
			
ج Init21	ج Init22	ج Init23	ج Init24
		Same as shape 11	
ج Init25	ج Init26	ج Init27	ج Init28

Note that the terminating shape of ب -init2 is very different from that of ج-init2. This is because they connect to different shapes of final 'bay'. This is discussed in the next section.

Final Forms:

With the exception of 'bay' and 'ray', all other letters have a unique final form. Both 'bay' and 'ray' have two final forms. These are shown in the table below. For 'bay', the first form occurs when it is preceded only by 'bay', (shape ب -init2 in the table above), 'fa', 'qaf', 'la' and 'ka', all in initial forms. While the other more frequent is realized else where and an example of letter connecting to it is 'jeem' having the form ج-init2. This explains the noticeable difference between the ending strokes of ب -init2 and ج-init2.

TABLE 7
FINAL FORMS FOR LETTER BAY AND RAY

	
ب Final 1	ب Final 2
	
ر Final 1	ر Final 2

The context for each of these shapes is given below. Interesting observation here is that 'ray-final1' occurs only when it is preceded by initial forms of 'bay' and 'jeem'. There is however no such precinct for letters 'ka' and 'la', which can be in any form (initial or medial).

ب → ب Final 1 / ب Initial Form ___
ب → ب Final 1 / ف Initial Form ___
ب → ب Final 1 / ق Initial Form ___
ب → ب Final 1 / ل Initial Form ___
ب → ب Final 1 / ک Initial Form ___
ب → ب Final 2 otherwise

ر → ر Final 1 / ب Initial Form ___
ر → ر Final 1 / ج Initial Form ___
ر → ر Final 1 / ک ___
ر → ر Final 1 / ل ___
ر → ر Final 2 otherwise

ACKNOWLEDGEMENT

The work presented has been partially supported by a Small Grants Program by IDRC, APDIP UNDP and APNIC. The work on Nastaliq has been completed by the Nafees Nastaliq team at CRULP, NUCES.

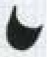
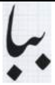
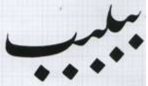
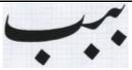
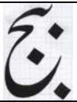
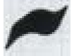
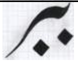
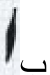
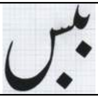
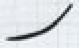
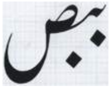
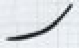
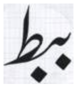
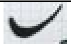
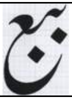
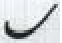
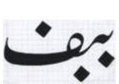
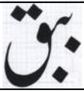
REFERENCE

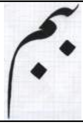
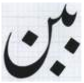
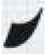
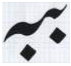
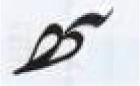
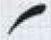
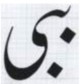
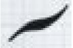
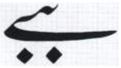
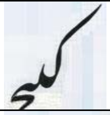
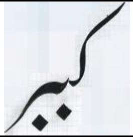
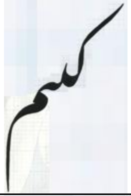
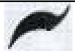
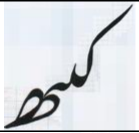
[1] www.ethnologue.com

[2] http://en.wikipedia.org/wiki/Aramaic_script

APPENDIX A: MEDIAL SHAPES OF BAY WITH EXAMPLES

Given below is a detailed analysis of 20 medial shapes and 4 initial shapes of letter ‘bay’. These 4 initial shapes are mentioned here since they are used in medial position of a ligature also. Following table lists context sensitive grammar of a particular shape and the shape’s glyph it self when it occurs in medial position. Also included is an example corresponding to each glyph.

ب → ب Medial1 / __<A> | __ ل | __ بInit2 | __ بInit6 | __ ب Init14 | __ ب Medial23 | __ ب Medial25	 بMedial1	
ب →ب Init2/ __ب Medi1 | __ ب Medi2 | __ ب Medi8 | __ ب Medi9 | __ ب Medi10 | __ ب Medi12 | __ ب Init15 | __ ب Medi16 | __ ب Medi18	 ب Init2	
ب → ب Medi2 / __ بFinal	 بMedi2	
ب → ب Medi3 / __ ج	 بMedi3	
ب → ب Medi5 / __ رFinal2	 بMedi5	
ب → ب Init6 / __ س	 ب Init6	
ب → ب Medi8 / __ ص	 ب Medi8	
ب → ب Medi9 / __ ط	 ب Medi9	
ب → ب Medi10 / __ ع	 ب Medi10	
ب→ ب Medi11 / __ ف	 ب Medi11	
ب → ب Medi12 / __ق __ و	 ب Medi12	

ب → ب Medi13 / __ م	 ب Medi13	
ب → ب Init14 / __ نFinal	 ب Init14	
ب → ب Init15 / __ ہ	 ب Init15	
ب → ب Medi16 / __ ھ	 ب Medi16	
ب → ب Medi17 / __ ی	 ب Medi17	
ب → ب Medi18 / __ ے	 ب Medi18	
ب → ب Medi23/ __ب Medi3	 ب Medi23	
ب → ب Medi25/ __ب Medi5	 ب Medi25	
ب → ب Medi33/ __ب Medi13	 ب Medi33	
ب → ب Medi36/ __ب Medi16	 ب Medi36	
ب → ب Medi37/ __ب Medi17	 ب Medi37