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".

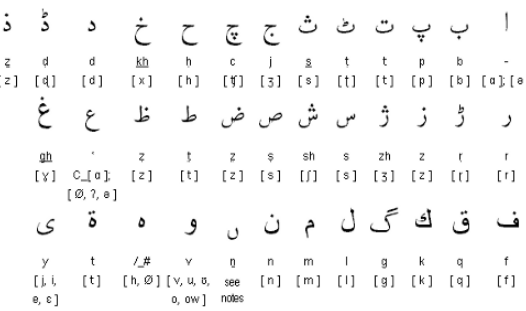
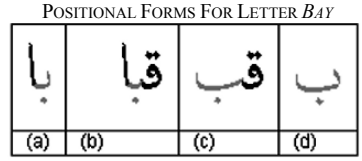


Fig. 1. Urdu Abjad

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*

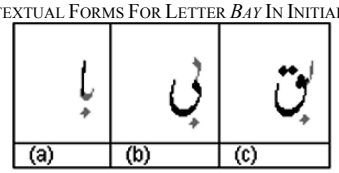


T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 53–58.

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.

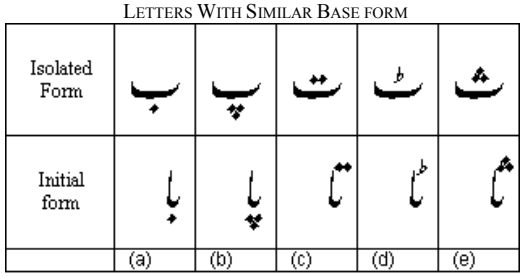
TABLE 2
CONTEXTUAL FORMS FOR LETTER *BAY* IN INITIAL FORM



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



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