FEATURES FOR NOORI NASTALIQUE

Aamir Wali, Atif Gulzar, Ayesha Zia, Muhammad Ahmad Ghazali, Muhammad Irfan Rafiq, Muhammad Saqib Niaz, Sara Hussain, and Sheraz Bashir

ABSTRACT

Most of the scripts existing today consist of huge inventory of characters. All characters have certain features that help perceive them differently from one another. Noori Nastalique script, like all other scripts, extracts certain characteristic features defined in both visual and articulatory terms. This paper uncovers these features and analyzes them.

1. INTRODUCTION

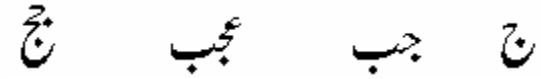
Till today lot of technological advancement have taken place in development of character recognition systems from pattern matching to the more dynamic feature extraction techniques. Developing such a system for Nastalique, a widely used Urdu font, by means of pattern or template matching would require a huge template inventory. This is because of context sensitive nature of Nastalique and according to previous research (Aamir..,2001) there are many as 474 shapes for 20 characters. Recognition through Feature extraction methodology can be an efficient solution to the Nastalique problem.

This paper analyzes and lists some features that can be employed to uniquely distinguish orthography of all alphabets and assist in their recognition. Note that all features are logical, that is they are features that a human mind will look for to perceive a character differently from other characters.

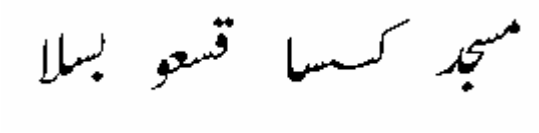
2. LITERATURE REVIEW

Nastalique is one of the most widely used Urdu fonts. Categorical division of written script is analogous to that used in phonology. For this reason we define corresponding concepts in similar way. Noori Nastalique is one of the most widely used Urdu fonts. Readers of this language systematically ignore certain properties of script and perceive two different shapes as the same character. We call the stored versions of written script as **graphemes**. Thus graphemes are the smallest unit of shape recognizable as an alphabet to the mind. That is, graphemes are how we mentally store different shapes of characters in our memory. All the different surface realizations of an underlying grapheme are its **allographs**. So such features should be devised that would satisfy all allographs of Nastalique.

Majority of the world's languages are unwritten (Fromkins, Victoria. 2000, p. 528). For most of the languages that are, their writing systems are simple and context free. Others are much more complex and actually have a context sensitive writing systems. Urdu is one such language. This complexity of Urdu is mainly due to couple of reasons. One is Urdu writing system is cursive. More than one character joins together to form a ligature. Important thing to be observed here is that the characters change their shape depending upon their position in the ligature. Each letter is written in a slightly different form depending on whether it comes in the beginning, middle or end of a word or whether it occurs on its own i.e. in a detached form. This is shown below:



In the above figure the character has formed four shapes according to its position. Another interesting property of Urdu writing system is that characters change their shapes depending upon the characters following and preceding it shown below. This change follows some rules like shape of the next letter to join with and the shape of the character, which is joining.



In the above figure it can be clearly seen that the second character *seen* (س) changes its shape in accordance with following and preceding characters. Thus, there can be multiple allographs of *seen* since we are mapping different visual image to the same grapheme.

A graphical unit can be parallel or complementary distributed depending on its environment. If two graphical units in the same environment have different meaning it implies parallel distribution as in case of graphemes. If environments for two graphical units are *mutually exclusive* it implies complementary distribution as in case of allographs.

The primary purpose of the present study was to examine all the characters of Noori Nastalique along with their allographs and devise features at logical level to describe and identify them.

4. Methodology

The basic study was done at a school (Z.N High School) on grade 2-4 students. The students were given some ligatures and asked to classify the characters in ligatures. When asked on how they distinguish characters, grade-2 students did not gave any justification. But grade 3-4 students could identify the characters mainly due to number and position of dots and due to some unique characteristics of some characters like, dot like head, number of dents etc.

The detailed study was done at CRULP with accordance to how children perceive characters and their various forms (allographs). The results of this analysis is stated and discussions next.

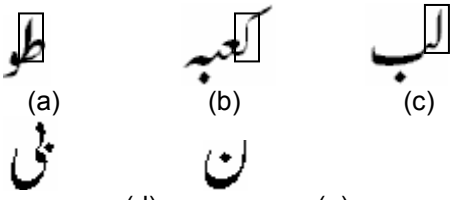
4. Results

Following is the list of features that can uniquely identify all the allophones identified in the previous section.

1. **Number of Dots:** Single or group of diacritics similar to a dot. This feature can have values 0,1,2,3.

2. **Position of Dots:** As can be seen these dots can be below or above a character body. A dot diacritic along a character body or below is considered to have the feature [-above].

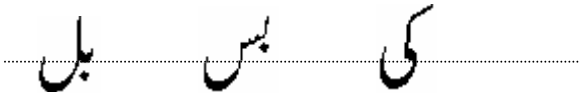
3. **Ascenders:** Prominent vertical stroke above the baseline. Below a, b, c are examples of ascenders, while d, e. are non-ascenders.



(a) (b) (c)

(d) (e)

4. **Descenders:** Feature describing an allograph visually prominent below a baseline.

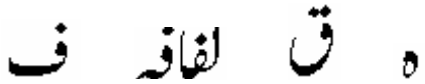


5. **Connected Forward:** Signifies if an allograph is connected to the following allograph. This can also be +ve or –ve. For all isolated and last position allographs (in a ligature) this value is negative. From the above feature ط and ب are positively connected forward.
6. **Kink:** sharp edges within an allograph.

7. **Diacritic:** Diacritics other than dots such as play a major role in allograph identification.

8. **Concavity:** Specifies the direction of a narrow opening within an allograph *above* the baseline. An allograph can have concavity left or right. Concavity upward or downward is not considered concave.

9. **Circular head:** This feature describes allographs that have a circular head. This may be filled or hollow.



10. **Ellipse:** This feature describes allographs that have a prominent elliptic-like shape.

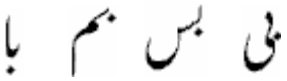
11. **Dot-like head:** A connected dot. In other words a dot forming a part of an allograph. This is not to be confused with a normal dot that is visibly separated from the allograph body.

12. **Diagonal:** A long prominent stroke inclined at a certain degree.

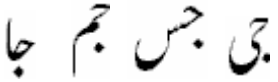
13. **Number of Dents:** A tooth like structure.

5. Discussion

In Urdu the diacritics play a major role in not only pronunciation but also in identification. Many of Urdu letters only differ by number of dots and their position. So first let us consider the features related to dot diacritics. Consider the following data where a dot below or along the character body is perceived as *bay*.



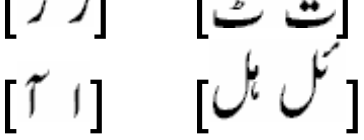
A dot at the same position can be confused with *jeem*:



But this does not happen. The kink containing head makes the difference. So we can describe *jeem* as +kink and *bay* as -kink.

Having said this, *khay* (similar to *jeem* but having a dot above) and *ghain* both have features [dots=1, +above, +kink]. This gave rise to another feature concavity already described in the previous section. *Khay* and *ghain* have the concavity left right respectively.

Coming back to the diacritics, consider the following pairs of data set.

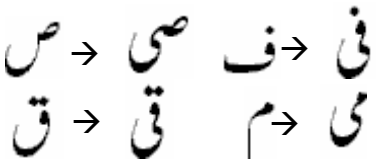


Clearly the character body in each pair is an exact duplication of one another, the only difference being their diacritics. They could therefore not be ignored and have been included in the list of features.

In Urdu, *lam* at the start or within a word look similar to *alif*. The cue to tell the difference is based on the knowledge that *alif can only occur in word final position* or isolated. Since *lam* looked like alif at start or in middle of the word, this led to the reason

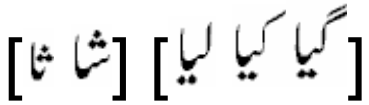
behind the feature connected forward. For *alif* it is [–connected forward], for *lam* this feature is [+connected forward]. *Lam* occurring at word final position goes well below the baseline, unlike *alif*. They were distinguished by the descenders property. *Alif* [-descender] and *lam* [+descender] both already have the [+ascender, -connected forward] features.

For some characters their identification depended entirely on orthography. Consider the following data, which shows how letters change shape from isolated to word initial form.



Based on these observations, some features were specified. On such feature was ellipse for the first example of the given data. Others were circular head and dot-like head for *fay and meem* respectively.

And finally turning to the diagonal and number of dents property. Consider the following data:



Can you tell what is the difference between the initial characters within respective brackets? In the first case the dent like structure of *sheen* makes it different from *say.* In the second case the diagonal is the only differing factor between the three ligatures. These two features were therefore included as there was no other way to answer the stated dissimilarities.

Following this discussion is a feature analysis for most of Urdu letters.

+ Ascender - Connected forward - Descender	ا
Number of dots = 1 -Above - Kink	ب
No. of dots = 1 -Above + Kink	ج
Number of Dents 3 Number of Dots = 0	س
Ellipse No of dots 0 +Descender	ص
Ellipse No of dots 0 -ascender	ه
Ellipse No of dots 0 + ascender	ط
No of dots = 0 +above Concavity = right + Kink + Descender	غ
Circular Head Number of Dots = 1 +above	ف
Circular Head Number of Dots = 2 +above	ق
Diagonal = 1 + Ascender	ک
Diagonal = 1 Circular Head	ک
Diagonal = 2 + Ascender	گ
Diagonal = 1 Circular Head	گ
+ Ascender + Descender - Connected Forward	ل
+ Ascender + Connected Forward	ل
Dot like head	م

Number of dots = 1 +Above - Kink	ن
Circular Head Number of Dots = 0	و
Ellipse Head Number o Dots = 0	ہ
Diacritic = Hook	ہ ھ
Ellipse = 2	ھ
Concavity = Right - Kink + Descender	ی
Concavity = Right - Connected Forward	ے

Some of the Graphemes, however could not be resolved using the features mentioned above are:

- *Dal* has a conflict with features of *ray* when they are connected with any other character.
- *Noon Gunnah* was unresolved or undefined for theses features.
- An allograph of *meem* cannot be defined with the features explained. This allophone occurs when it is connected forward with *alif*.
- Two allophones of *gol hay* are to be handled as exceptions since they are used in a particular context. They are ﷲ and ۃ.

6. References

1. *Azhar-ul-Lughat Urdu.*
2. Fromkins, Victoria A. 2000. *Linguistics: An Introduction to Linguistic Theory.* Blackwell Publishers inc. Oxford, UK
3. Aamir, Atif, Ayesha, Ahmad, Irfan, Saqib, Sara and Sheraz. "Contextual Shape Analysis of Nastaliq", CRULP Annual Report 2001.

Appendix A Graphemes of Urdu