CONTEXTUAL SHAPE ANALYSIS OF NASTALIQ

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

ABSTRACT

Nastaliq calligraphic style is one of the most complex and widely used styles of Urdu script. It employs different shapes and sizes for the same letter in a bewildering variety of contexts. It is observed that these different shapes vary with neighboring letters as well as position of that letter in a ligature. This paper explores this variety in shapes that Nastalique offers, thus providing a foundation to eventually model this inherent variety and recreate this script font as written by a calligrapher.

1. INTRODUCTION

Nastaliq is one of the most widely used fonts of Urdu script in Pakistan and is regarded as one of the most complex fonts in the literature of electronic computing. Early efforts to make this font available electronically used the approach of storing all possible ligatures (sequence of Urdu characters occurring without space) of Urdu, which was not supported by the standard Microsoft Windows font technology of TTF (True Type Font) available at that time. Consequently this font was limited to some specialized Urdu Word Processors only. With the emergence of new font technology OTF (Open Type Font) from Microsoft, it is now possible to make Nastaliq font, which would be portable to any standard word processor like Microsoft Word, Power Point and Excel.

A natural approach for this, is to make a character-based Nastaliq font i.e. in spite of storing all possible ligatures, individual characters are stored. Character-based Nastaliq font needs to specify the rules, which govern the change of shape of characters while moving from one segment of ligature to another. The first step of such a complex writing system is to identify the all-possible shapes of characters in ligatures and then study the rules that change the shape in particular context.

The aim of this paper is to identify all possible shapes of characters based upon the context in which they occur in ligatures, thus providing a foundation to eventually model it.

2. LITERATURE REVIEW AND PROBLEM STATEMENT

2.1 Urdu Writing System

Urdu is a complete language with its own script, which is a mixture of Arabic and Persian script. Urdu script has total 38 alphabets excluding Aerabs (Vowel Marks). Character set and Aerabs of Urdu are shown below in Fig 1.

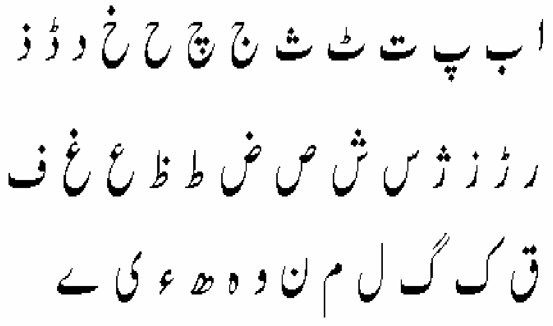


Figure 1(a) Character Set of Urdu

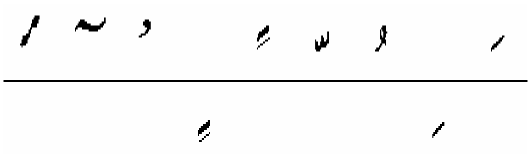


Figure 1(b) Aerabs of Urdu

From the above figure it can be observed that several of the basic alphabets of Urdu share same shape, these are differentiated only by the placement of dots or diacritic Tuay on the basic shape. This property makes Urdu script fast and easy to write.

Urdu is written in the opposite direction to English i.e. from right to left. An interesting concept about Urdu is that its number system is written form left to right. So Urdu writing system has both the properties of left to right and right to left writing systems as shown in figure 2.

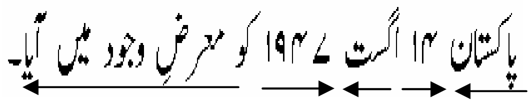


Figure 2 Urdu Number System

Urdu writing system is cursive. More than one character joins together to form a ligature. Important thing to be observed 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 in isolated form. This is shown in the figure 3.

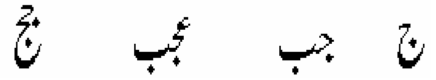


Figure 3 Different shapes according to the position in ligature

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 as in figure 4. This change follows some rules like, shape of the next letter to join with and the shape of the character, which is joining.

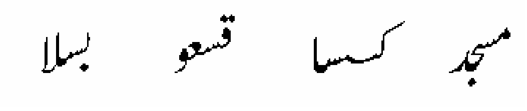


Figure 4 Different Shapes of the Letter Seen (س).

In the above figure it can be clearly seen that the second character (س) changes its shape in accordance with following and preceding characters. Which symbolizes that Urdu writing system is context sensitive.

2.2 Different Fonts of Urdu

Urdu fonts started growing after the Khat-e-Kofi, an Arabic font that is now being used by Urdu. It was invented in the year 238 AH. But now Kofi has lost its fame because of its complex writing system as shown in figure 5.

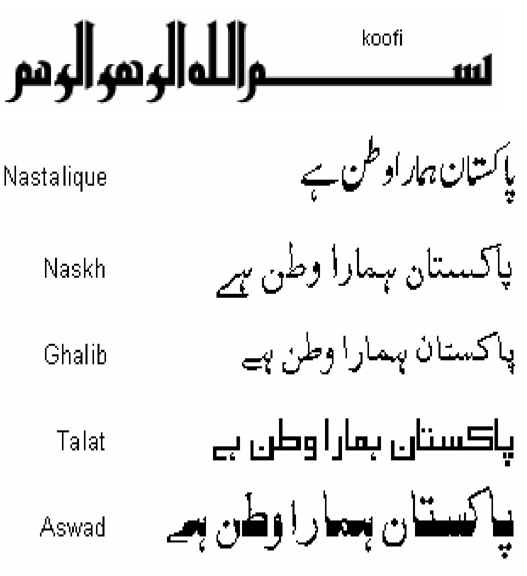


Figure 5 Different fonts of Urdu

In the year 310 AH a famous Muslim scholar Ibne-Muqalla invented a new and easy font called Naskh. All the previously invented fonts lost their prominence due to Naskh's easy and fast writing system.

Although a considerable number of studies on Urdu Naskh script have been conducted during the last few years and consequently led to the development of software related to Urdu Naskh, studies on Urdu Nastaliq are still in their premature stages. This is mainly due to extraordinary features of this unique and dynamic form of writing. There are many other fonts of Urdu, some of which are shown in the figure 5.

2.3 Urdu Nastaliq Script

Urdu Nastaliq script is a collection of two other scripts Naskh and Talique. These two scripts were combined to form another script formally called Naskh-Talique, which was shortened to Nastaliq. Hence, this 38 alphabets script has properties of both Naskh and Talique.

Two most common feature of Nastaliq found in Naskh or for that matter in any Persian or Arabic script is that it is cursive. That is, the tip of the pen is not raised until a ligature is complete. Another characteristic is that Nastaliq is written from right to left unlike English which is form left to right. In addition to these, there are other characteristics of Nastaliq that have made its automation difficult.

Nastaliq is actually written from top right to bottom left. Each ligature is tilted at approx. 45 degree (See Fig. 6). This is of particular significance as there is no fixed level or height for any character. Positions of characters change at a continuous scale. In Naskh each character has four shapes depending on whether the character is isolated, at start, in middle or coming at end. In Nastaliq, however characters may have significant amount of variation at one position alone. This is because of the influence of other characters in the same ligature.

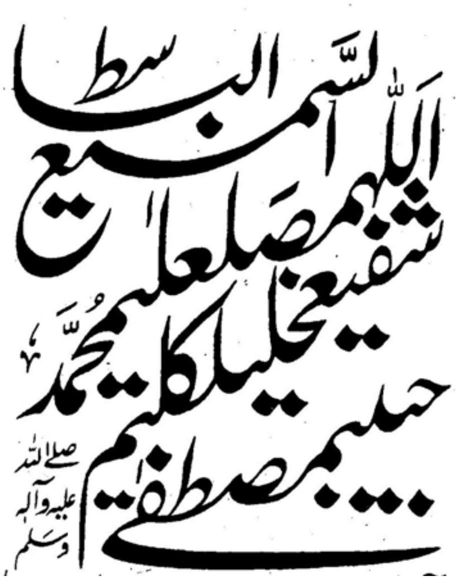


Figure 6 Nastaliq diagonality

The aim of this paper is to identify all possible shapes of characters of Nastaliq style based upon the context in which they occurs in ligature, thus providing a foundation to eventually model it.

Noori Nastaliq (a type of Nastaliq font, see fig. 5) provides an ideal platform for this kind of analysis. Moving from right to left this style of writing uses simple spacing rules (where size of a shape does not depend on how much space is available for the letter). Thus Noori Nastaliq style of writing eliminates other complexities of Nastaliq font while providing different shapes and their context.

3. METHODOLOGY

We need to study all possible ligatures of Noori Nastaliq, to identify different shapes from them. Practically speaking, it is not possible to study all of the ligatures. So we limit our study up to 4 character ligatures (which are about 600,000 ligatures).

We observed that some of them behave similarly under all circumstance, while writing. Which can be observed from the following example in figure 7. In this figure the middle characters have common shapes with the difference of dots.

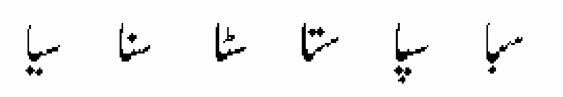
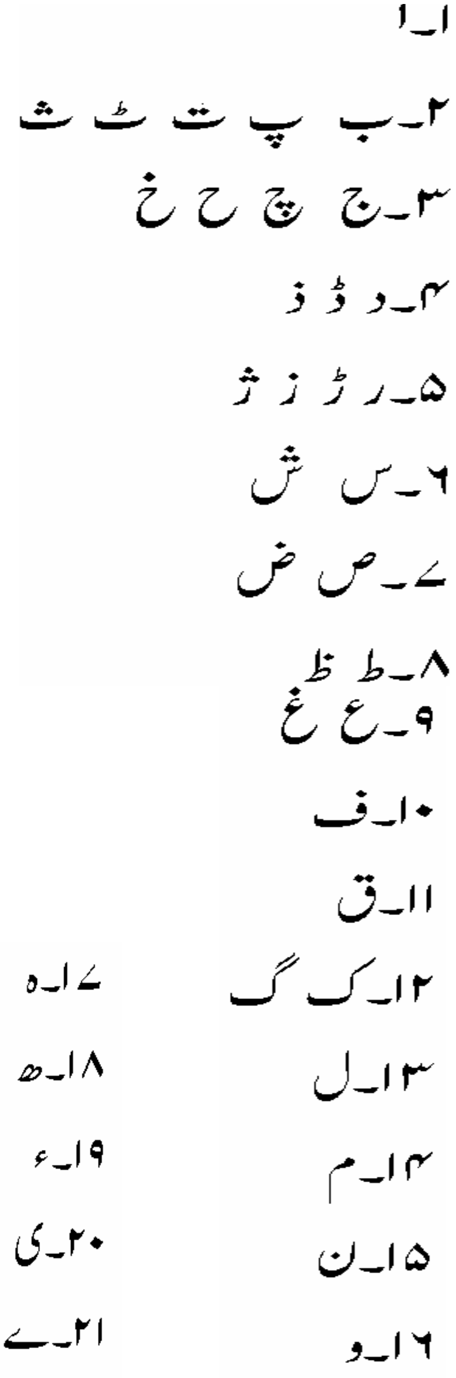


Figure 7 Examples of characters behaving similarly in all the contexts

So these 38 characters have been categorized into 21 classes so that any member of same class will have same shape in ligature provided the same context This minimal subset is shown in Table 1.

Table 1: Classification of Urdu characters



۱۔ ا

۲۔ ب پ ت ٹ ث

۳۔ ج چ ح خ

۴۔ د ڈ ذ

۵۔ ر ڑ ز ژ

۶۔ س ش

۷۔ ص ض

۸۔ ط ظ

۹۔ ع غ

۱۰۔ ف

۱۱۔ ق

۱۲۔ ک گ

۱۳۔ ل

۱۴۔ م

۱۵۔ ن

۱۶۔ و

۱۷۔ ہ

۱۸۔ ھ

۱۹۔ ء

۲۰۔ ی

۲۱۔ ے

The next step was to identify all possible shapes of all categories. For this purpose a well-known Urdu word processor Inpage version 1.1.34 was used. Typing of large number of combination was a cumbersome and an impossible job. So, a simple program was written that could generate all possible combinations of 2 characters, 3 characters and 4 characters ligature and store them in a text file.

The combinations of 2 to 4 characters were chosen because most of the character in a ligature changes it shape depending upon the shape of the subsequent two to three characters. This can be better understood from the figure 7.

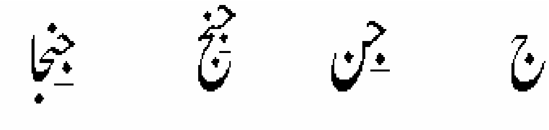


Figure 7 Change in the shape of a character depending upon the context

In the above figure the character Geem (ج) is in its isolated form. When some other character follows it, it changes its shape, which proves the change in shape according to the next character in 2 characters ligature. Similarly this property can be observed for the 3 and 4 character ligatures.

The text files generated by the program were then imported into the Urdu software so that they can be treated as Urdu characters. We carefully observed all ligatures containing a particular character from the minimal subset that we previously determined (Table 1). Any variation in shape was noted and recorded.

4. RESULTS

The following table shows the list of Urdu letters along with the number of shapes that were found for that letter. Appendix A contains the details of the shapes that were found.

No. of Shapes	Letter	No. of Shapes	Letter
64	ق	2	ا
44	ک	52	ب
29	ل	37	ج
31	م	3	د
52	ن	3	ر
1	و	33	س
34	ہ	37	ص
28	ھ	22	ط
52	ی	50	ع
52	ے	64	ف

5. DISCUSSION

One major factor that affects the accuracy of above results is the technique of 'Perceptual examination' i.e. the decision that two shapes in different ligatures are same or not, are identified by human perception.

The other source of inaccuracy could be the fact that Inpage has stored all the ligatures of Noori Nastaliq written by hand. So we might have identified some additional shapes caused by the variation of hand written.

The collected data has enabled us to carry another study to investigate the rules that might have caused the change in shape of these characters. This will help us to better understand the nature of Nastaliq.

7. REFERENCES

Ethnologue: Languages of the World, 14th Edition, 2001 © Summer Institute of Linguistics, URL: I http://www.sil.org

Appendix A

اردو نستعلیق کے حروف تہجی کی تمام ممکن اشکال

2	:	ا
52	:	ب
37	:	ج
3	:	د
3	:	ر
33	:	س
37	:	ص
22	:	ط
50	:	ع
64	:	ف
*1	:	ق
44	:	ک
29	:	ل
31	:	م
*1	:	ن
1	:	و
34	:	ہ
28	:	ھ
*1	:	ی
*1	:	ے

*ب،ن،ے،ی کی درمیان اور شروع مین ایک ہی شکل ہوتی ہے .اسی طرح ف،ق کی بھی ایک ہی شکل ہوتی ہے۔

Total : 474

ا

۱۔ا ۲۔با

ب

۱۔با ۲۔بب ۳۔بج ۴۔بر ۵۔بس

۶۔بص ۷۔بط ۸۔بع ۹۔بف ۱۰۔بق

۱۱۔بم ۱۲۔بن ۱۳۔بھ ۱۴۔بہ ۱۵۔بی

۱۶۔بے ۱۷۔ببا ۱۸۔ببر ۱۹۔بہن ۲۰۔ببے

۲۱۔بعب ۲۲۔بعق ۲۳۔بنج ۲۴۔بکب ۲۵۔بہا

۲۶۔بب ۲۷۔کب ۲۸۔جیب ۲۹۔ببج ۳۰۔ببر

۳۱۔ببہ ۳۲۔لبی ۳۳۔ببے ۳۴۔ببس ۳۵۔نبیق

۳۶۔ببکا ۳۷۔ببکب ۳۸۔سیند ۳۹۔سنبل ۴۰۔طبیب

۴۱۔طیبہ ۴۲۔طبقا ۴۳۔طبقہ ۴۴۔عبیر ۴۵۔عینک

۴۶۔کینا ۴۷۔لینی ۴۸۔لینے ۴۹۔منبع ۵۰۔بسبس

۵۱۔کلبد ۵۲۔کطبم

ج

۱۔جا ۲۔جب ۳۔بج ۴۔جر ۵۔جس

۶۔جص ۷۔جع ۸۔جف ۹۔جم ۱۰۔جن

۱۱۔جو ۱۲۔جج ۱۳۔جبر ۱۴۔جبک ۱۵۔جبی

۱۶۔جسے ۱۷۔جفا ۱۸۔جلب ۱۹۔جکل ۲۰۔جہا

۲۱۔جہب ۲۲۔جھا ۲۳۔جیب ۲۴۔جیل ۲۵۔جنجا

۲۶۔بج ۲۷۔بجا ۲۸۔بجب ۲۹۔بجج ۳۰۔بجر

۳۱۔بجس ۳۲۔بجص ۳۳۔بجع ۳۴۔بجہ ۳۵۔ججا

۳۶۔بجبر ۳۷۔بجلب

د

۱۔د ۲۔بد ۳۔سد

ر

۱۔ر ۲۔بر ۳۔سر

س

۱۔سا ۲۔سب ۳۔سج ۴۔سر ۵۔سس

۶۔سص ۷۔سط ۸۔سع ۹۔سم ۱۰۔سن

۱۱۔سو ۱۲۔سہ ۱۳۔سی ۱۴۔سے ۱۵۔سبب

۱۶۔سبن ۱۷۔سبج ۱۸۔سبد ۱۹۔سطی ۲۰۔سعف

۲۱۔سکب ۲۲۔سلک ۲۳۔سبر ۲۴۔سنب ۲۵۔سہ

۲۶۔نسفق ۲۷۔سہس ۲۸۔بسعو ۲۹۔جبجج ۳۰۔سلسم

۳۱۔سہا ۳۲۔بمسر ۳۳۔منمس

ص

۱۔صا ۲۔صب ۳۔ص ۴۔صد ۵۔صر

۶۔صق ۸۔صم ۸۔صو ۹۔ص ۱۰۔صبا

۱۱۔صبر ۱۲۔صبس ۱۳۔صبی ۱۴۔صص ۱۵۔مصل

۱۶۔صصا ۱۷۔صطر ۱۸۔صعا ۱۹۔صعب ۲۰۔صکا

۲۱۔صلب ۲۲۔صنے ۲۳۔صہا ۲۴۔صھا ۲۵۔صید

۲۶۔بصا ۲۷۔بصب ۲۸۔بصج ۲۹۔بصر ۳۰۔صم

۳۱۔بصو ۳۲۔بصہ ۳۳۔بصی ۳۴۔بصھ ۳۵۔سصو

۳۶۔فصو ۳۷۔لصّف

ط

۱۔طا ۲۔طب ۳۔طج ۴۔طر ۵۔طس

۶۔طص ۷۔طع ۸۔طف ۹۔طق ۱۰۔طم

۱۱۔طہ ۱۲۔طی ۱۳۔طے ۱۴۔طبا ۱۵۔طبر

۱۶۔طبق ۱۷۔طبو ۱۸۔طبہ ۱۹۔طبی ۲۰۔طمب

۲۱۔طہا ۲۲۔بط

ع

۱۔عا ۲۔عب ۳۔عج ۴۔عر ۵۔عس

۶۔عص ۷۔عع ۸۔عم ۹۔عو ۱۰۔عی

۱۱۔عے ۱۲۔عبا ۱۳۔عبج ۱۴۔عبد ۱۵۔عبر

۱۶۔عجب ۱۷۔عصا ۱۸۔عفا ۱۹۔بع ۲۰۔بجا

۲۱۔بعج ۲۲۔بعس ۲۳۔بعہ ۲۴۔بعے ۲۵۔جعد

۲۶۔جعو ۲۷۔سعو ۲۸۔صعب ۲۹۔صعر ۳۰۔صعو

۳۱۔طعن ۳۲۔فعا ۳۳۔معر ۳۴۔نعو ۳۵۔بعبر

۳۶۔بعکب ۳۷۔سعید ۳۸۔طعنو ۳۹۔طعنے ۴۰۔کعبہ

۴۱۔کعبے ۴۲۔لعبا ۴۳۔لعفو ۴۴۔لعیب ۴۵۔معبو

۴۶۔معجم ۴۷۔معمو ۴۸۔نعبد ۴۹۔نعلو ۵۰۔بجعی

ف

۱۔فا ۲۔فب ۳۔فج ۴۔فد ۵۔فر

۶۔فص ۷۔فص ۸۔فع ۹۔فف ۱۰۔فق

۱۱۔فک ۱۲۔فل ۱۳۔فم ۱۴۔فن ۱۵۔فو

۱۶۔فی ۱۷۔فے ۱۸۔فہ ۱۹۔فبب ۲۰۔فبر

۲۱۔فبس ۲۲۔فجو ۲۳۔فجب ۲۴۔فجم ۲۵۔فجر

۲۶۔فجھ ۲۷۔فسا ۲۸۔فصو ۲۹۔فصے ۳۰۔فصا

۳۱۔فطر ۳۲۔فعب ۳۳۔فکا ۳۴۔فکل ۳۵۔فکس

۳۶۔فلف ۳۷۔فما ۳۸۔فمس ۳۹۔فنا ۴۰۔فنس

۴۱۔فہص ۴۲۔فہا ۴۳۔فیف ۴۴۔لبف ۴۵۔بفا

۴۶۔بفج ۴۷۔بفر ۴۸۔بفس ۴۹۔بفص ۵۰۔بفق

۵۱۔بفبا ۵۲۔بفبب ۵۳۔بفبج ۵۴۔بفبر ۵۵۔بفبس

۵۶۔بفبو ۵۷۔بفبے ۵۸۔بفجا ۵۹۔بفجب ۶۰۔بفجج

۶۱۔بففہ ۶۲۔بفکا ۶۳۔بفہ ۶۴۔جفنا

ک

۱۔کا ۲۔کب ۳۔کج ۴۔کد ۵۔کر

۶۔کس ۷۔کص ۸۔کع ۹۔کک ۱۰۔کم

۱۱۔کن ۱۲۔کو ۱۳۔کہ ۱۴۔کھ ۱۵۔کے

۱۶۔کبر ۱۷۔کبس ۱۸۔کبج ۱۹۔کبی ۲۰۔کبے

۲۱۔کجد ۲۲۔کجب ۲۳۔کجک ۲۴۔کجہ ۲۵۔کسس

۲۶۔کعج ۲۷۔کعر ۲۸۔کفا ۲۹۔کقا ۳۰۔ککا

۳۱۔ککے ۳۲۔ککر ۳۳۔ککو ۳۴۔کمب ۳۵۔کنہ

۳۶۔کہا ۳۷۔کہب ۳۸۔کھا ۳۹۔کھل ۴۰۔کھج

۴۱۔کھص ۴۲۔رکا ۴۳۔بباک ۴۴۔بمک

ل

۱۔ل ۲۔لا ۳۔لب ۴۔لج ۵۔لد

۶۔لر ۷۔لس ۸۔لص ۹۔لط ۱۰۔لع

۱۱۔لف ۱۲۔لق ۱۳۔لک ۱۴۔لل ۱۵۔لم

۱۶۔لن ۱۷۔لھ ۱۸۔لہ ۱۹۔لی ۲۰۔لجد

۲۱۔لسس ۲۲۔لکم ۲۳۔للّٰہ ۲۴۔بلب ۲۵۔للّٰه

۲۶۔بلکل ۲۷۔بلھب ۲۸۔بلھ (~۱۷)۲۹۔بلھب (~۲۷)

(~اس شکل میں ل چھوٹی ہوتی ہے۔)

م

۱۔ما ۲۔مب ۳۔مج ۴۔مر ۵۔مس

۶۔مص ۷۔مع ۸۔من ۹۔مہ ۱۰۔می

۱۱۔مے ۱۲۔مبد ۱۳۔مبج ۱۴۔مسی ۱۵۔مصا

۱۶۔مطا ۱۷۔مطر ۱۸۔مکب ۱۹۔ملب ۲۰۔مما

۲۱۔مم ۲۲۔جمع ۲۳۔جمن ۲۴۔جمو ۲۵۔جمی

۲۶۔ جے ۲۷۔ جمر ۲۸۔ سما ۲۹۔ سمس ۳۰۔ سمن
۳۱۔ ہمم

و

۱۔ و

ہ

۱۔ ہ ۲۔ ہا ۳۔ ہب ۴۔ ہج ۵۔ ہر
۶۔ ہس ۷۔ ہص ۸۔ ہع ۹۔ ہف ۱۰۔ ہن
۱۱۔ ہھ ۱۲۔ ہہ ۱۳۔ ہج ۱۴۔ ہبد ۱۵۔ ہجد
۱۶۔ ہکا ۱۷۔ ہ کل ۱۸۔ ہہا ۱۹۔ بہا ۲۰۔ بہب
۲۱۔ ہج ۲۲۔ بہر ۲۳۔ بہس ۲۴۔ بہص ۲۵۔ بہع
۲۶۔ بہف ۲۷۔ بہق ۲۸۔ بہم ۲۹۔ بہن ۳۰۔ بہو
۳۱۔ بہھ ۳۲۔ بہہ ۳۳۔ بہ ۳۴۔ للہ

ھ

۱۔ھا ۲۔ھب ۳۔ھج ۴۔ھر ۵۔ھس

۶۔ھص ۷۔ھق ۸۔ھم ۹۔ھن ۱۰۔ھو

۱۱۔ھ ۱۲۔ھھ ۱۳۔ھی ۱۴۔ھے ۱۵۔ھبا

۱۶۔ھبب ۱۷۔ھبج ۱۸۔ھبر ۱۹۔ھبو ۲۰۔ھبے

۲۱۔ھجا ۲۲۔ھجب ۲۳۔ھفھ ۲۴۔ھکل ۲۵۔ھما

۲۶۔ھھا ۲۷۔بباھ ۲۸۔بپھ