L2/08-343

Dr. Attash Durrani*

GHOST CHARACTERS:
Atomization- Combination Theory

Phoenicians gave us the letters to write or record our messages in ideographic form. Before this innovation there were atleast two styles of writing systems: (1) Cuniforms of Summarians: using arrows for writing and (2) Hieroglyphic of Egyptians: using pictures to depict messages. The Phoenicians develped some basic signs: a conversion of pictorial elements into outlines to depict phonemes. These outlines were converted into two types of scripts (1) Arabic and (2) Roman, emerging from the same roots. CJK ideographs are another different story.

In Arabic script there are 29 letters derived from the Phonecian alphabet e.g.:-

ء ا ب ج د ھ و ز ح ط ی ک ل م ن س ع ف ص ق ر ش ت ث خ ذ ض ظ غ

Initially the Arabs were not using dots and diacritics to write the letters with these ideogrphs. The dotless arabic script was depicting basic letters like the following:

ء ا ٮ ح د ھ و ر (ح) ط ی ک ل م ں
س ع ڡ ص ٯ (ر س ٮ ٮ ح د ص ط ع)

It means that they were using only 19 ghost characters and reading these letters by their cultural habit of usage and connotations. It were the non-arabs who could not become at home with these ghost characters or "Khali Kashtian" (Dotless

*** Project Director, Center of Excellence for Urdu Informatics, National Language Authority, Islamabad***

letters), so why Hajaj Bin Yousuf introduced dots to differenciate different letters emerging from the same roots of ideographs: one, two or three dots above or below.

The ghost (Khali) characters or common shapes of letters, we may call them "Kashties", were 19 in number. Holy Quran was inscribed in these Khali characters. The letters of the Holy Prophet (PBUH) were also written in these Kashties. The Arabs had no difficulty to read unwritten atoms (dots and diacritics) to utter the sound of any letter. These 19 characters were as follows:

1. ء
2. ا،أ،آ=ا
3. بتث = ٮ
4. جحخ=ح
5. دذ=د
6. رز=ر
7. سش=س
8. صض=ص
9. طظ=ط
10. عغ=ع
11. ف = ڡ
12. ق= ٯ
13. ک=ک
14. ل=ل

15. م=م

16. ن=ں

17. ه=ه

18. و وٰ وَ=و

19. ییٰیَ=ی

The philosophy behind to dot the ghost characters, as coined by Ibne Maqla, was this that the first letter will have one dot, then two or three dots; first letter will have lower dots and then upper, may be introduced like:

بتث

جحخ

دذ

رز

س ش

صض

طظ

عغ

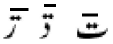
فق

ںن

Persia adopted Arabic script after the expansion of Islamic teachings and added some letters having three dots or one more line e.g.: پ(ب) چ(ج) ژ (ز) گ(ک)

When this script came into use in India, some Hindi

Urdu sounds (Phonemes) were to be derived adding four nuqtas in the ghost characters like: ٿ ڐ ڙ

These again converted the diacratics and dots having two dots and one line like: 

After a long calligraphic practice the final shape of these letters in Urdu bacame in shape like this: ٹڈڑ

The Hindi phonemes or diphthongs were depicted with "ہ" like بہ،پہ،تہ،ٹہجہ،چہ،دہ،ڈہ،رہ،ڑہ،کہ،گہ

The final shape is now using "ھ" character shape like:بھ، پھ، تھ، ٹھ، جھ، چھ، دھ، ڈھ، رھ، ڑھ، کھ، گھ.

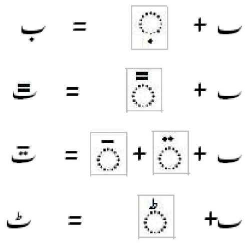
Another "Yeh" "ی" was written in Urdu with half "ی" and then was converted into "ے".

It means that in Persian and Urdu languages three basic ghost characters were added, e.g. گ،ھ،ے

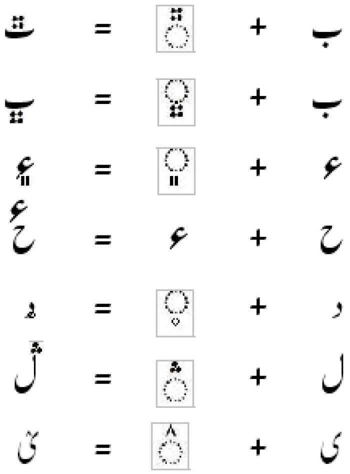
A total number of 22 ghost characters are now in use in all the languages using arabic script. These are as follows:

اٮ ح د ر س ص ط ع ڡ ٯ ک گ ل م ں و ہ ء ی ھ ے

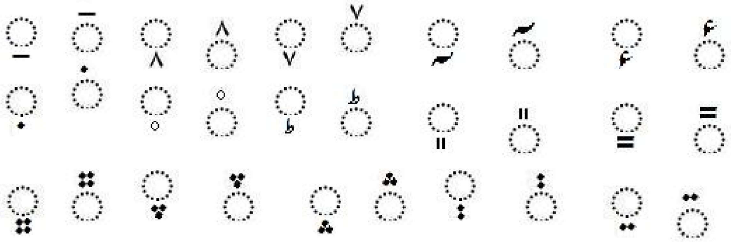
Any letter of any language written in Arabic script has only 22 ghost characters in its basic system. The dots (Nuqtas), dandas, Toey and other diacritics or atoms are added for a combination or normalization to form any letter like:



This is the basic theme of Atomization or combination. All other languages like Sindhi, Pashto, Balti, Balochi, Shina etc. use the same technique to develop their letters e.g.:-



It means that there are atleast 30 dots and diacratics as atoms that combine to the basic ghost characters to form a letter in Arabic script to be used by any language. These are in upper and lower placements as follows:



A total of 52 (22 + 30) atoms or characters form at least 660 different letters. These 660 can add one more ghost atom from dots and diacritics to form 19800 formal letters. The possibility of combining atoms to form letters goes up to a number of 20460 (600+ 19800 = 20460). It is a futuristic approach. If some one needs to see all these letters, it will require 80 pages of UNICODE having 256 letters on one page. Some of these possible 20460 letters may be seen in the annex.

As an initial step some atoms of Urdu letters were placed in the ASCII code plate version 2 of NLA prepared in 1999 and exhibited in ITCN Asia 2001, Karachi., UNICODE Consortium took this initiative inclusive of Bey, Fey, Qaf without dots to place them in its version 3.1 and in published version 4.00 in Arabic Block but it ignored dots and other diacritics. A proposal was sent to the UNICODE consortium on 29-1-2006 to add dots or nuqta characters. Some 22 new combining Characters were to be added in the Arabic Block of Unicode. The addition of these combining marks will not only simplify the Arabic Block of Unicode Standard, but will make it possible to typeset all origional languages written in Arabic script. It was said a complimentary addition. These were:

1. Arabic Single Nuqta Above
2. Arabic Single Nuqta Below
3. Arabic Double Nuqta Above.
4. Arabic Double Nuqta Below
5.Arabic Tripple Nuqta Above
6. Arabic Tripple Nuqta Below
7. Arabic Tripple Inverted Nuqta Above
8. Arabic Tripple Inverted Nuqta Below
9. Sindhi Quadrple Nuqta Above
10. Sindhi Quadrple Nuqta Below
11. Sindhi Double Danda Above
12. Sindhi Double Danda Below
13. Sindhi Double Nuqta Vertical Above
14. Sindhi Double Nuqta Vertical Below
15. Urdu Single Kashida Above
16. Urdu Single Kashida Below
17. Urdu Double Kashida Above
18. Urdu Double Kashida Below
19. Pashto Single Circle Above

20. Pashto Single Circle Below
21. Urdu Letter Tota Above
22. Urdu Letter Tota Below

This way, not only we are able to support all the characters of all languages in Arabic script, but even if some appear later on, we are already future compatible.

Even today, from pedagogical point of view, the different characters are memorized by teaching their basic shape and the number of Nuqtas with their orientation.

The inclusion of these combining characters to the Arabic Block will simplify the representation of texts in Arabic scripts. Not only existing characters can be defined as a character sequences but new characters also can be formed which are even not yet encoded into the standard.

The Arabic script is mostly being used by the societies which are evolving and hence their scripts are open ended. New characters are being defined on a rapid pace for many regional languages but encoding all of them to the Unicode Standard is inconvenient. Such an approach will merely stuff the Arabic Code Block very rapidly and will require extensions to the block on almost yearly basis.

The presence of U+066E ARABIC LETTER DOT LESS BEH, U+066F ARABIC LETTER DOT LESS QAF and U+06A 1 ARABIC LETTER DOT LESS FEH is useless in the absence of these combining characters.

It is requested to provide code points for the above mentioned combining characters to simplify the process of Arabic Script processing.

Below some references are given. The first reference is about 300 years old (Muqalaat-e-Hafiz Mehmood Sheerani) and is of historic nature. Other important reference is the dictionary published by Anjuman-e- Taraqqi-e-Urdu (Pakistan) in 1985, which was edited by the late Molvi Abdul Haq (also known as Father of Urdu, Baba-e-Urdu) (2).

Farhang-e- Talaffuz is another important document published by National Language Authority (Government of Pakistan) (3).

Sheena Quaida is an educational text published by Sarhad Provincial Government (Government of Pakistan) with support from Himalay Jungle Project financed by European Community and British High Commission Islamabad (4) and Balti Quaida published by National Language Authority (5).

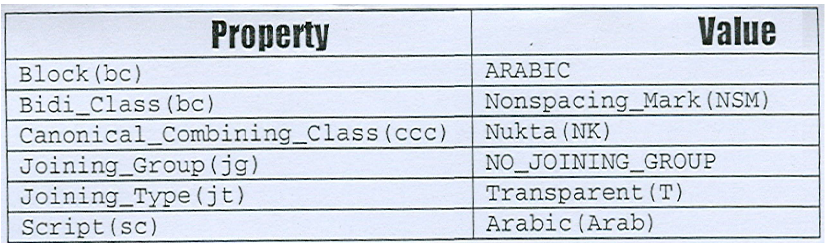
Moreover, Government of Pakistan is developing a standard Nastaleeq font usable for all the regional language. This font is based on the idea of Character Composition and hence to be practical, needs support of Nuqta in the Unicode Standard.

Detailed and formal properties of these characters will be provided after the initial response from the UTC.

The Unicode technical committee was presented with a proposal (document number L2/06-039) to add Nuqta marks to the Unicode Arabic Block on 6th February 2006. That document contained the characters to be added and their rational for addition to the Arabic Block.   .

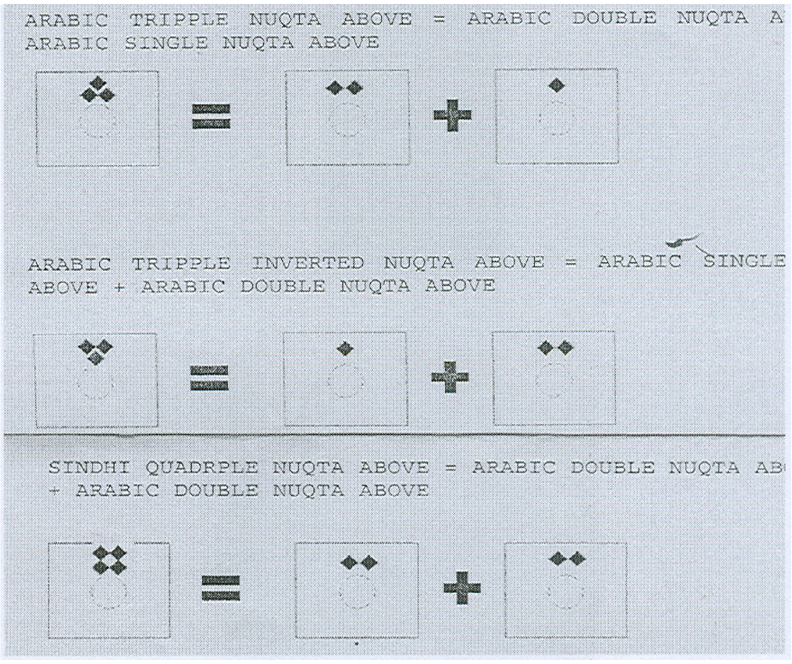
The intent of this document was to provide further detailed character properties and elaborate their status, usage and advantage of their presence in the Arabic Block.

Almost all the proposed characters posses same character properties so for the sake of simplicity, all are listed in the table below:



Property	Value
Block(bc)	ARABIC
Bidi_Class(bc)	Nonspacing_Mark(NSM)
Canonical_Combining_Class(ccc)	Nukta(NK)
Joining_Group(jg)	NO_JOINING_GROUP
Joining_Type(jt)	Transparent(T)
Script(sc)	Arabic(Arab)

Further more, Nuqta characters having further canontical decompositions are list along with there canonical decompositions. We only present the above Nuqta can case for below Nuqta is Indentical to be above one.



ARABIC TRIPPLE NUQTA ABOVE = ARABIC DOUBLE NUQTA A
ARABIC SINGLE NUQTA ABOVE

ARABIC TRIPPLE INVERTED NUQTA ABOVE = ARABIC SINGLE
ABOVE + ARABIC DOUBLE NUQTA ABOVE

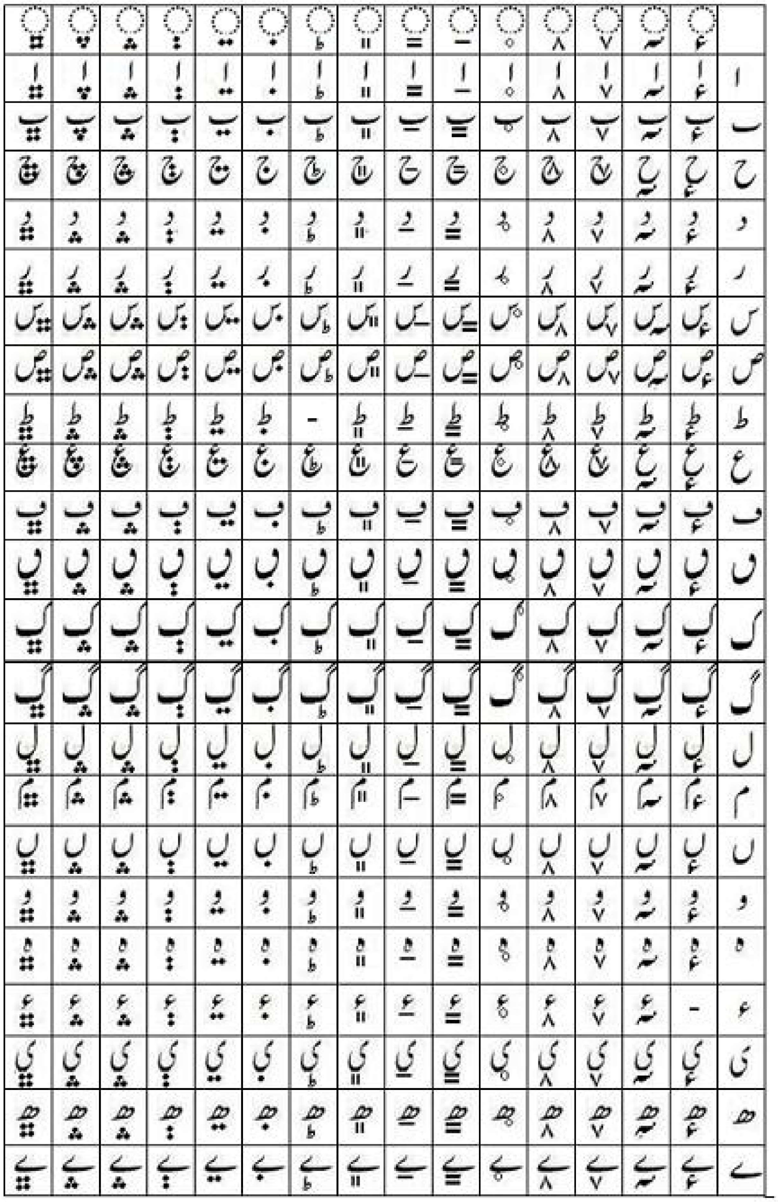
SINDHI QUADRPLE NUQTA ABOVE = ARABIC DOUBLE NUQTA AB
+ ARABIC DOUBLE NUQTA ABOVE

The atomization -combination rules were already present in the UNICODE, but without any theoretical elaboration that is now formed to achieve a good model for Arabic Script. This helps to develop a chart of atom combinations for character formation. Next are the Single Combination and Double Comibnation charts. There may be

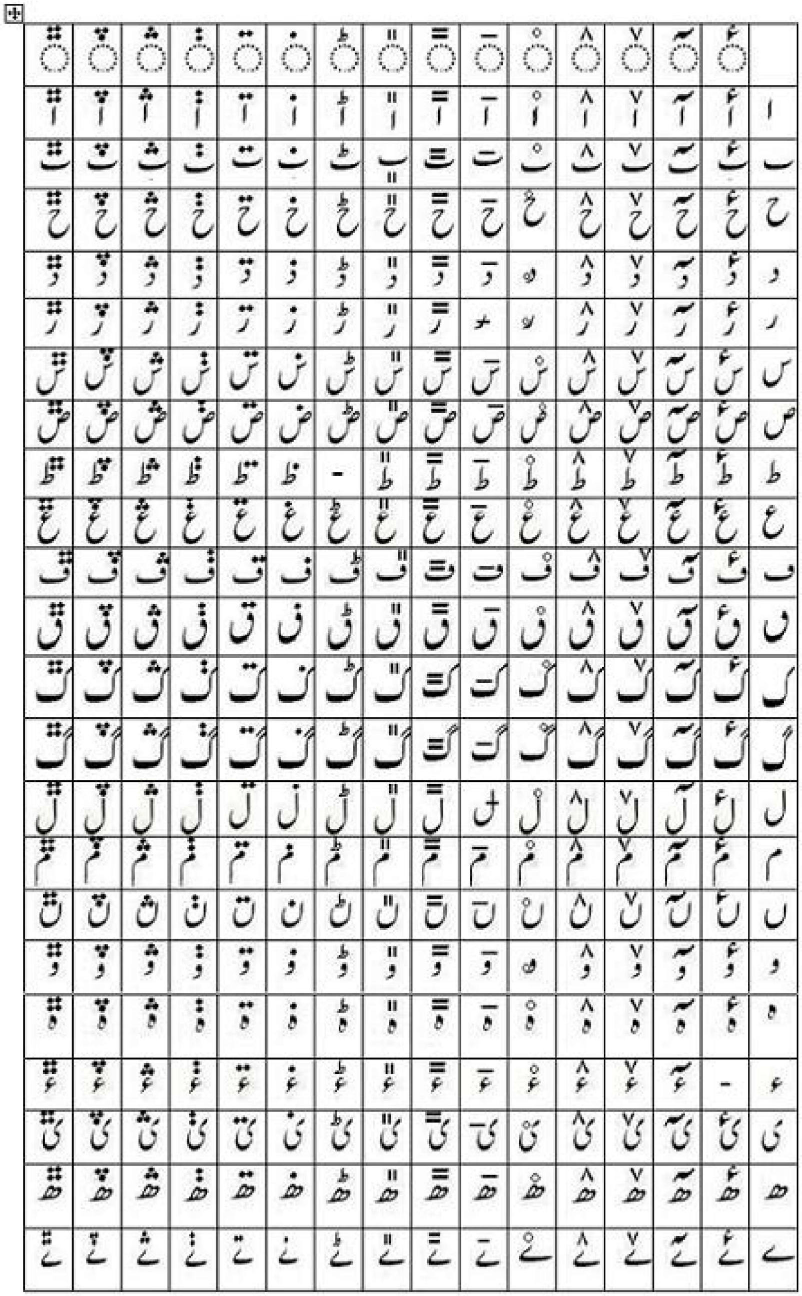
660 letter formations of single combination depicted in next two pages and out of 20460 possible letter formation only those characters are mentioned which are already on Unicode.

One can easily find that the Ghost Charaters Theory saves the place on International Standards, i.e. (22+30) 52 spaces and 20460 possible letters are made to happen from past to far future of the Arabic Script to be used by any language of the world or for any pedogogical teaching purposes.

Single Combination Chart
(Below Nuqta Characters)


ا
ٮ
ح
د
ر
س
ص
ط
ع
ڡ
ٯ
ک
گ
ل
م
ں
و
ہ
ء
ی
ھ
ے

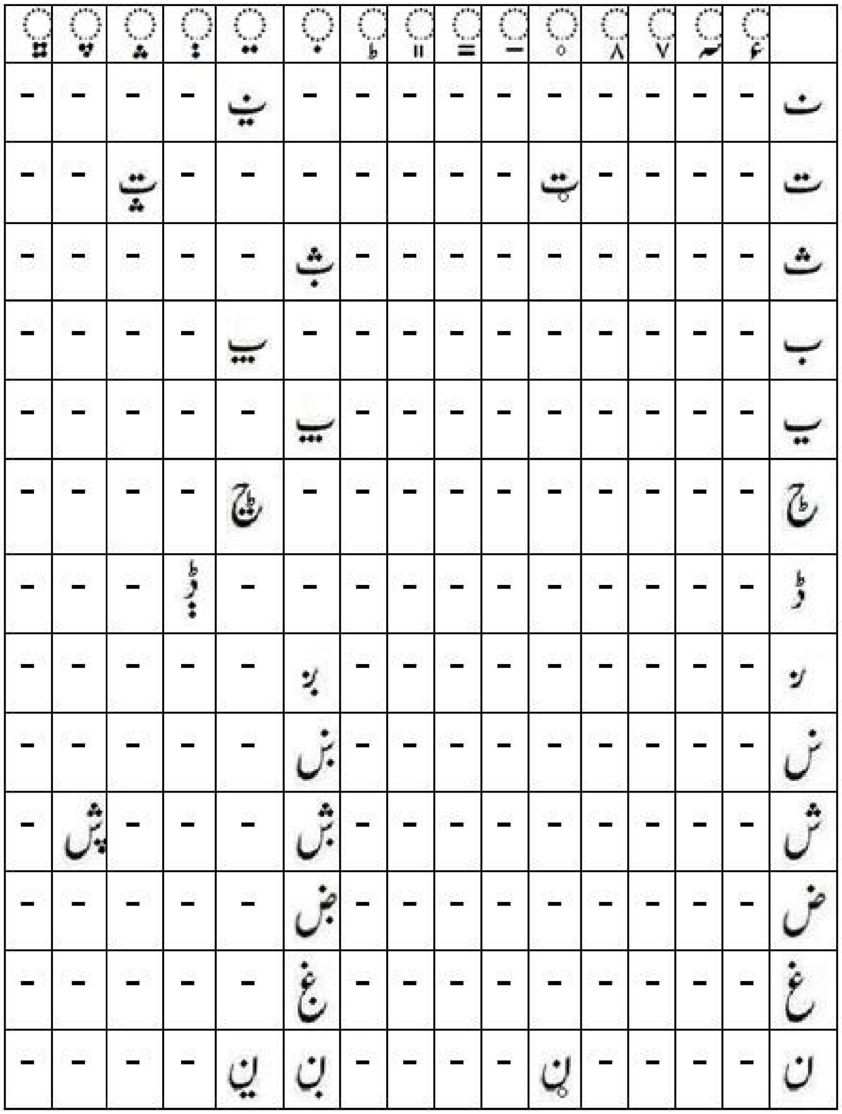
Single Combination Chart
(Upper Nuqta Characters)



Double Combinations Unicode Examples (Upper Nuqta Characters)

◌∷	◌∵	◌ۛ	◌:	◌̈	◌̇	◌ؕ	◌̎	◌̿	◌̄	◌۟	◌ٛ	◌ٚ	◌ٓ	◌ٔ	
-	-	-	-	ٮ̥̈	-	-	-	-	-	-	-	-	-	-	ٮ̥
-	-	بۛ	-	-	-	-	-	-	-	-	-	-	-	-	ب
-	-	-	-	-	ٻ̇	-	-	-	-	-	-	-	-	-	ٻ
-	-	-	-	پ̈	-	-	-	-	-	-	-	-	-		پ
-	-	-	-	-	-	ر̤ؕ	-	-	-	-	-	-	-	-	ر̤
-	-	-	-	-	ڔ̇	-	-	-	-	-	-	-	-	-	ڔ
	-	-	-	-	-	ڗؕ	-	-	-	-	-	-	-	-	ڗ
-	-	بۛں	-	-	ب̇ں	-	-	-	-	-	-	-	-	-	بں
-	-	پۛں	-	-	-	-	-	-	-	-	-	-	-	-	پں
-	-	-	-	-	-	تؕں	-	-	-	-	-	-	-	-	تں
-	-	-	-	-	ض	-	-	-	-	-	-	-	-	-	ص
-	-	-	-	-	غ	-	-	-	-	-	-	-	-	-	ع
-	-	-	-	-	ل̥̇	-	-		-	-	-	-	-	-	ل̥
-	-	-	-	-	ل̣̇	-	-	-	-	-	-	-	-	-	ل̣
-	-	-	-	-	ل̤̇	-	-	-	-	-	-	-	-	-	ل̤
-	-	-	-	-	-	نؕ	-	-	-	-	-	نٚ	-	-	ن

Double Combinations Unicode Examples
(Up/Below Nuqta Characters)



◌﮻	◌﮹	◌﮷	◌﮾	◌﮵	◌﮳	◌﯁	◌﮼	◌=	◌ـ	◌﮿	◌Λ	◌V	◌~	◌ٕ	
-	-	-	-	ٮ﮲﮵	-	-	-	-	-	-	-	-	-	-	ٮ﮲
-	-	ت﮷	-	-	-	-	-	-	-	ت﮿	-	-	-	-	ت
-	-	-	-	-	ث﮳	-	-	-	-	-	-	-	-	-	ث
-	-	-	-	ب﮵	-	-	-	-	-	-	-	-	-	-	ب
-	-	-	-	-	ٻ﮳	-	-	-	-	-	-	-	-	-	ٻ
-	-	-	-	ݯ	-	-	-	-	-	-	-	-	-	-	ݮ
-	-	-	ڑ﮾	-	-	-	-	-	-	-	-	-	-	-	ڑ
-	-	-	-	-	ږ	-	-	-	-	-	-	-	-	-	ز
-	-	-	-	-	س﮲﮳	-	-	-	-	-	-	-	-	-	س﮲
-	ش﮹	-	-	-	ش﮳	-	-	-	-	-	-	-	-	-	ش
-	-	-	-	-	ض﮳	-	-	-	-	-	-	-	-	-	ض
-	-	-	-	-	غ﮳	-	-	-	-	-	-	-	-	-	غ
-	-	-	-	ن﮵	ن﮳	-	-	-	-	ن﮿	-	-	-	-	ن

Bibliography:

1. Muqaalaat-e-Hafiz Mehmood Sheerani, No.VIII, (ed. Mazhar Mahmood Sheerani), Published by Majlis Tarraqi-i-Adab, Lahore.

2. The Standard Urdu-English Dictionary by BaBa-e-Urdu Molvi Abdul-Haq, published by Anjman-e-Traqqi-e-Urdu, 1985.

3. Farhang-e-Talaffuz by Shaan-ul-Haq Haqqi published by National Language Authority, Pakistan, 1995.

4. Sheena Qaida, publised by N.W.F.P Government, Pakistan (2004).

5. Balti Qaida, published by National Language Authority of Pakistan, 2004.

6. Jonathan Kew, Encoding Arabic extensions: options for the future of Unicode, Unicode: Document L2/03/004, February 11, 2003.

7. The Unicode Standard, (version) 5.0, The Unicode Consortium, published by Addison-Wesley, Upper Saddle River, NJ, October 2006.

8. Tanveer Fatima, Ghost Characters Theory, nlauit.gov.pk.

9. Urdu Informatics (Vol, II), (ed. Dr. Attash Durrani), published by National Language Authority (Pakistan), 2008.

10. nlauit.gov.pk/e.magazine