Proposal of Inclusion of Certain Characters in Unicode

Muhammad Numan Chishti
Iqbal Academy Pakistan
sa@iap.gov.pk

Abstract

Two characters (marks) of Urdu are part of Urdu Zabita Takhti (UZT 1.01) but not present in Unicode (5.1.0). There is at least one mark and one character of Urdu, written in books and dictionaries, which is neither part of UZT, nor of Unicode. After reviewing several dictionaries, it is observed that without inclusion of these three symbols, development of a correct dictionary and a text-to-speech system for Urdu is not possible. This paper recommends the inclusion of these four in Unicode.

1. Introduction

Urdu is native language of 182 million people; it has more than 270 million speakers and is fourth most spoken language (after Chinese, English and Spanish) in the world[1]. The language has a history of more than 700 years. But its character set was not standardized till January 26, 2004[2]. However, the variants of these characters as well as the marks are not standardized by any body or authority yet. No complete, (almost error free) dictionary, corpus or lexicon of Urdu with correct marks has yet been prepared using Unicode. Some serious efforts, including a few sponsored, have been made for preparation of computerized Urdu dictionaries, however, none is comparable with any standard/printed one. Among several causes of the flaw, one is absence of the certain marks that are used for producing proper pronunciation. Similarly, text-to-speech systems cannot produce correct sound till they are trained with correct marks for any sound.

2. Scope

This paper is limited to inclusion of three symbols i.e. (i) “Leta Pesh” (UZT # Hex-47) or “Arabic Damma Majhool”; (ii) “Leti Zer” (UZT # 48) or “Kasra Majhool” and; (iii) “Alif-e-Wavi”, representation of “Noon Ghunna” (U+06BA, UZT# 71) when appears in middle form, and modifying definitions of (i) “Arabic Zwarakay” (U+0659) for its use as “Leta Zabar” or “Fatha Majhool” and; (ii) “Arabic Vowel Sign Inverted Small V Above” (U+065B) for its use as “Ulta Jazm”. There is a possibility of absence of other characters too.

3. Revision of UZT

Urdu Zabita Takhti Version 1.01 (Urdu Code Page) was standardized by the Government of Pakistan in July, 2001 whereas the character set of Urdu was standardized by the National Language Authority in January, 2004. Thus, the UZT may need a revision in the light of the approved character set.

4. Dictionary entries

A typical dictionary entry includes these parts:

1. The word or phrase broken into syllables.
2. The word or phrase with the pronunciation indicated through the use of diacritical marks - marks that indicate the vowel sounds such as a long vowel or a vowel affected by other sounds; accent marks, a mark called the schwa that tells you that the vowel is in an unaccented syllable of the word.
3. the part or parts of speech the word functions as -for example as a noun (n.), verb (v.), adjective (adj.), or adverb (adv.).
4. Related forms of the word, such as the plural form of nouns and the past tense of verbs.
5. The definition or definitions of the word or phrase. Generally dictionaries group the definitions according to a word's use as a noun, verb, adjective, and/or adverb.
6. The origin, or etymology, of the word or words, such as from the Latin, Old French, Middle English, Hebrew, the name of a person. Some dictionaries use the symbol < to mean "came from." For example, the origin of

the word flank is given as "<Old French *flanc*<Germanic."This tells us that *flank* came from the Old French word *fanc*. The French word in turn came from the German language. Some dictionaries use abbreviations to tell you where the item came from: OE for Old English, L for Latin, and so forth.

4.1. Webster's NewWorld Dictionary

In Webster's NewWorld Dictionary of American English[3] (Third Edition), the syntax of entries is as follows:

Main entry word (pronunciation) **part-of-speech label. Inflected form**, <The Etymology>, The Definition, ***USAGE*** Labels & Notes

For example:

bazaar (bə zär') ***n.*** [Pers *bāzār*, a market] **1** a market or street of shops and stalls, esp. in Middle Eastern countirs **2** a shop for selling various kinds of goods **3** a sale of various articles, usually to raise money for a club, church, etc.

4.2. Oxford Advanced Learner's Dictionary

In Oxford Dictionary[4], the syntax of an entry is:

headword (also **alternative spellings of headword)**/pronunciation/part of speech/definition

For Example:

Bazaar/bə'za:(r)/n **1** (in eastern countries) group of shops or stalls or parth of a town where these are. **2** (in Britain, USA, etc) (place where there is a) sale of goods to raise money for charitable purposes: *a church bazaar.*

4.3. فرہنگِ تلفظ

In Farhang-e-Talaffuz[5], the syntax of an entry is as follows:

Main entry word[6]. part-of-speech label**.** The Definition

For Example:

بازارا، مذ۔ وہ جگہ جہاں خرید و فروخت ہو، جہاں بہت سی دکانیں ہوں ہاٹ، پینٹھ، تجارتی مال کی کھپت کا حلقہ، مارکیٹ، منڈی، شارع عام، گزرگاہِ عام، ج مذ، **بازارگاں**: بیوپاری، تاجر لوگ، نیز بازارگان صفت: **بازاری:** بازار سے متعلق، عام، گھٹیا، معمولی، ناشائستہ، عصمت فروش۔

Bāzār. Ism. mūẓkkar. Wo jgha jhāṉ khrīd o frokht ho, jhāṉ buht sī dukānyṉ huṉ. Haṭ, pyṉṭ, tjarti māl kī khpt kā ḥalqa, markīt, manḍi, shar'a 'ām, guzrgah i 'ām, jam'a mūẓkkar, **Bāzārgaṉ**: byopari, tajir log, nīz Bāzārgan ṣift: **Bāzārī**: Bāzār kay mut'aliq, 'ām, gheṭiya, m'amūlī, nā shaisth, 'īṣmt frwsh.

Thus, it is clear that marks are integral part of all dictionaries.

5. Marks (اعراب) in Urdu:

There are following marks in Urdu:

Name	Unicode	Unicode Description
Alef Maksura: ا	U+0627	Arabic letter Alef
Alef Mamduda: آ	U+0622	Arabic letter Alef with Madda above ≡ u+0627 + u+0653
Alef Bala ٰ	U+0670	Arabic letter superscript Alef
Alef Gher malfuz أ	U+0627 + U+06EB	Arabic letter Alef with Arabic center high stop
Alef-e-Wavi ٓٔ	U+0622 + U+06EB	Arabic letter Alef with Madda above with Arabic center high stop above Madda
Alef Zerin ٖ	U+0656[9]	Arabic subscript Alef
Waw Mar'uf[10] ؤ	U+0648 + U+0657	Arabic letter waw with Arabic inverted Damma on it.
Waw Majhool ۥ	U+0648 + U+0659	Arabic letter waw with Arabic zwarakay[11]
Waw Lain وَ	U+064E + U+0648	Arabic letter waw with Arabic Fatha on previous character

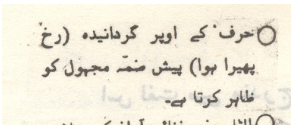
Waw Madulah وٓ (Waw Gher malfuz)[12]	U+0648 + U+06EB	Arabic letter waw with Arabic center high stop
Yah-e-Mar'uf ی	U+0650 + U+06CC	Arabic Letter Farsi Yeh with Arabic Kasra on previous character
Yah-e-Majhool ۓ	U+6D2 + {Kasra Majhool}	Arabic Letter Yeh Barree with Arabic Kasra majhool
Yah-e- Lain ے	U+064E + U+06D2	Arabic Letter Yeh Barree with Arabic Fatha on previous character
Fatha (Zabar) ـَ	U+064E	Arabic Fatha
Kasra (Zer) ـِ	U+0650	Arabic Kasra
Damma (Pesh) ـُ	U+064F	Arabic Damma
Fatha Majhool (Leta Zabar) ٙ	U+0659	Arabic Zwarakay
Kasra Majhool (Leti Zer) ٖ	Inclusion in Unicode is proposed in this paper	
Damma Majhool (Leta Pesh) ٚ	Inclusion in Unicode is proposed in this paper	
Damma M'akus (Ulta Pesh) ٗ	U+0657	Arabic Inverted Damma
Fatha Maghnoona ـَنْ	U+064E + U+0646 + U+065A	Arabic letter Noon with Arabic vowel sign small V above and Arabic Fatha on previous character
Kasra Maghnoona ـِنْ	U+0650 + U+0646 + U+065A	Arabic letter Noon with Arabic vowel sign small V above and Arabic Kasra on previous character
Damma Maghnoona ـُنْ	U+064F + U+0646 + U+065A	Arabic letter Noon with Arabic vowel sign small V above and Arabic Damma on previous character
Noon Sahi ن	U+0646	Arabic Letter Noon
Noon Ghunna ں	U+06BA	Arabic Letter Noon Ghunna
Meem bshakal-e-Noon ںٛ	U+0646 + U+065B	Arabic Letter Noon with Arabic vowel sign inverted small V above
Tashdeed ّ	U+0651	Arabic Shadda
Jazm ْ	U+06E1	Arabic Jazm
Ulta Jazm ٜ	U+065A	Arabic vowel sign small V above
Fatahtan (Do Zabar) ً	U+064B	Arabic Fatahtan
Kasratan (Do Zer) ٍ	U+064D	Arabic Kasratan
Dammatan (Do Pesh) ٌ	U+064C	Arabic Dammatan

6. Definition of "Majhool"

The lexical meaning of "Majhool" is little unknown. It is used for the sound which is not available in a greater language. In the context of this paper, the word majhool is used for the sound which is unknown or not available in Arabic.

7. Damma Majhool (Leta Pesh) ٚ

The symbol of Damma Majhool (Leata Pesh) is a small circle accompanied by a horizontal line towards its right at the top of any character.

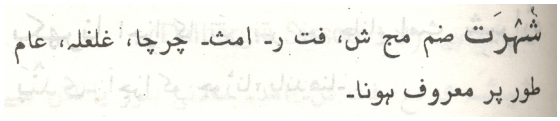


(1) حرف کے اوپر گردانیدہ رُخ پھیرا ہوا پیش، ضمّہ مجہول کو ظاہر کرتا ہے۔[13]

Ḥarf ke ūpar gardānydah rukh phayra huya pesh, ẕmma majhūl ko ẓahir kerta hay.

"Averted *pesh* over a character represents *ẕmma majhūl*."

This is in of the group which is present in UZT (Hex-47) but not recommended by Dr. Khawar Zia for inclusion into Unicode. Examples of its use are given below:-



(1) شہرت: ضم مج ش، فت ر۔ امث، چرچا، غلغلہ، عام طورپر معروف ہونا۔

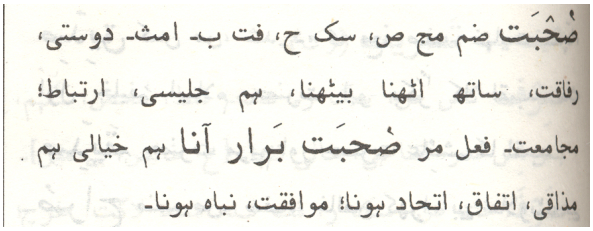
Shuhrat: zmma majhūl shīn, fatha ra, ism muanus, charchā, ghulghulha, 'ām ṭur par marūf hona.

"Fame: Averted *pesh* on *shīn*, tilted line above *ra*. Noun feminine. Admiration, tumult, known to common."

(2) کہرام: ضم مج ک، سک ہ، امذ، رونے دھونے کا شور، واویلا.

Kuhrām: zmma majhūl kāf, skūn ha, ism mūẕkkar, ronay dhonay ka shur, wawayla.

"Lamentation: Averted *pesh* on *kāf*, amputation on *ha*, Noun masculine. Noise of weeping, crying."



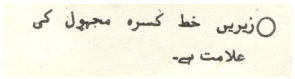
(3) صحبت ضم مج ص، سک ح، فت ب۔ امث۔ دوستی، رفاقت، ساتھ اٹھنا بیٹھنا، ہم جلیسی، ارتباط؛ مجامعت۔

Ṣuḥbat zmma majhūl sād, skūn ha, fatha ba. ism muanus. Dosti, rfaqt, sath ūṭhna bayṭhna, ham jlaysi, irtbaṭ; mujam'at

Company: Averted *pesh* on *sād*, amputation on *ha*, Noun masculine. Friendship, closeness, living together, acquaintance, association

8. Kasra Majhool (Leti Zer) ◌ِ

The symbol of the mark Leti Zer is a small line under the character.



زیریں خط کسرہ مجہول کی علامت ہے.

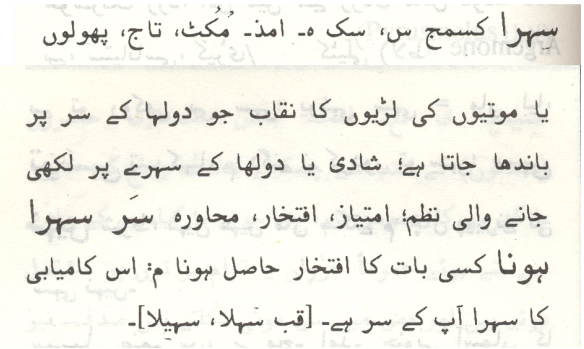
Zayrīn khat kasrah majhūl kī 'alamt hay.

"Underline of a character is the symbol of *kasrah majhūl*."

Kasra Majhool (or Leti Zer) appeared in UZT (HEX 48) in 2001. Unicode version 3.2 was available at that time. In last seven years, Unicode has gone through several additions including two major updates in Unicode 4.0 and 5.0. However, this mark has not yet been included.

Dr. Khaver Zia, in his presentation "Towards Unicode Standard for Urdu" delivered during First National Urdu Software Development Workshop held in March, 2001 at FAST-NU, Lahore grouped this symbol in a category of characters of UZT which are not part of Unicode (3.2) but inclusion into Unicode was not recommended by him.

It is observed that Kasra Majhool (Leti Zer) is widely used in Urdu marks to indicate proper pronunciation of certain words. Here are a few examples[14]:-

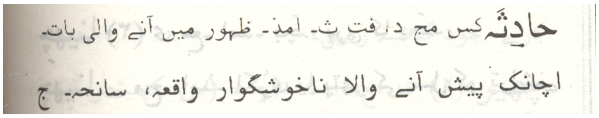


(1) سہرا:۔ سہرا: کس مج س، سک ہ۔ امذ۔ مُکٹ، تاج، پھولوں یا موتیوں کی لڑیوں کا نقاب جو دولہا کے سر پر باندھا

جاتا ہے، شادی یا دولھا کے سہرے پر لکھی جانے والی نظم، امتیاز، افتخار۔

Sayhra:- kasrah majhūl sīn, skūn ha, ism mūz̤kkar. Tāj, phūluṇ ya motīṇ kī lṛyūṇ ka nīqab ju dulha kay sar par baṇdha jātā hay, shādi yā dulha kay sayhray par līkhī janay walī naz̤m, imtīāz, iftikhār.

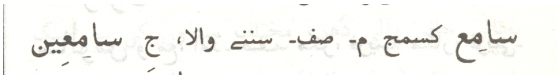
"Chaplet: underline on *sīn*, amputation on *ha*, Noun masculine. Tiara, crown, hood of string of flowers or pearls which is being tied on the head of groom. A poem written on the occasion of wreathing a groom or on marriage, distinction, elegance."



(2) حادثہ: کس مج د، فت ث۔ امذ۔ ظہور میں آنے والی بات۔ اچانک پیش آنے والا ناخوشگوار واقعہ، سانحہ۔

Ḥādṡh: kasrah majhūl dāl, fatha ṣa, ism mūz̤kkar. Z̤hūr maiṇ āānay valī bāt. Achānk pesh āānay valā nā khushgvār vaq'a, sānḥā.

"Accident: underline on *dāl*, tiled line above *ṣa*, Noun masculine. Occurrence of news, occurrence of an unfortunate happening, mishap."



(3) سامع: کس مج م۔ صف۔ سننے والا

Sam'e: kasrah majhūl mīm. Ṣift, sunnay wala

"Listener: underline on *mīm. Adjective.* One who hears."

9. Alif-e-Wavi آ

Though, its use is very limited, It can be defined as:- Arabic letter Alef with Madda above with Arabic center high stop above Madda. However, Shan ul Haq Haqqui and Molvi Abdul Haq have used this notation to produce sound of ô for writing words like Ball or Call in Urdu. As, it is neither بال (bāl), nor بول (bol) or کال (bāl) or کول (kol). The sound is in between of the two.

یہ علامت انگریزی الفاظ کے لیے مخصُوص ہے بہ ضمن اشتقاق م Call, Ball۔

Yah 'alamt aṇgrayzī alfāz̤ kay līay makhṣūṣ ha bhz̤mn ishtīqāq masāl call, ball.

"This symbol is reserved for some English words, for example: call, ball."

10. Noon Ghunna ں

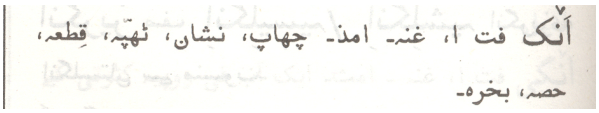
Noon Ghunna when appears in isolated form, it appears as "ں" (U+06BA). However, when it appears in the center of a word, like چونکہ, موند، سنگ، تنگ مانگ its correct glyph is a small "v" above "ن". This symbol is widely used in all the dictionaries as well as several literary books. However, it is absent from Unicode.

"ن کی دو حالتیں ہوتی ہیں۔ ایک تو جب اِس کی آوازپوری ادا ہو جیسے "پان"، "گیان"، "دھیان" وغیرہ۔ دوسرے جب پورے طورپر نہ ادا ہو بلکہ کسی قدرناک میں گنگنی سی آوازنکلے، ایسی حالت میں اسے نون غُنہ کہتے ہیں۔ جیسے سماں، کنواں، سانپ، اینٹ، ہنسنا وغیرہ میں۔ نون غُنہ جب آخر میں آتا ہے اس میں نقطہ نہیں دیتے لیکن بیچ میں آتا ہے تو اس ہر اُلٹا جزم لگانا چاہیے (v)"[۱۵]

Nūn kī dw ḥāltaiṇ hotī hayṇ. Aayk tw jb īs kī āāvāz purī adā ho jaysey pān, gayan, dhayan, wgayrha. Dūsray jb puray ṭur par na adā ho blkh kīsī qdr nāk maiṇ gungunī sī āāvāz nīlay, aysī ḥālt maiṇ usay nūn ghunha kehtay hayṇ. Jaysey samāṇ, kuṇwāṇ, sāṇp, īṇt, hṇsnā wgayrha maiṇ. nūn ghunha jb āākhr maiṇ āātā ha īs maiṇ nūqta nahīṇ datay laykin bīch maiṇ āātā ha tw us pr ūlṭā jzm lgāna chāhīay.

"There are two styles of *nūn*. First when it gives its full sound like *pān*, *gayan*, *dhayan*, etc. Second when it delivers partial sound from nose. In such a situation it is called *nūn ghunha*. For example: *samāṇ*, *kuṇwāṇ*, *sāṇp*, *īṇt*, *hṇsnā* etc. When *nūn ghunha* appears at the end of a word, the dot inside *nūn* is not placed. However, when it appears in-between a word, an inverted *jzm* is placed on top of the dot."

Almost all the authentic dictionaries have used this correct symbol for Noon Ghunna. For Example:-



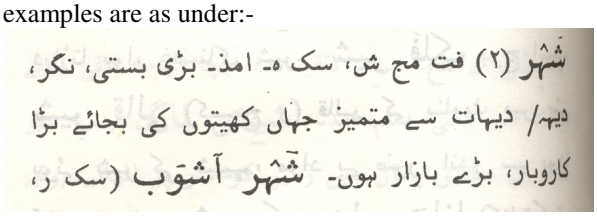
اَنک: فت ا، غنہ۔ امذ۔ چھاپ، نشان، ٹھپہ، قطعہ، حصہ، بخرہ۔

Āṇk: fatha Alef, ghunha. ism mūẕkkar. Chāp, Nishān, Ṭhappa, Qiṭh, ḥissh, bkhrh.

"Impression: tiled line above *Alef*, ghunha. Noun masculine. Stamp, Mark, Inkling, impact."

11. Fatha Majhool (Leta Zabar) ◌ُ

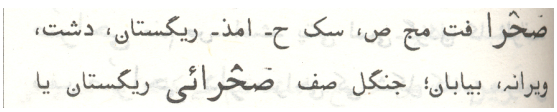
A horizontal line above a character represents Fatha Majhool. Its examples include شہر (city), صحرا (dessert). It is neither شَہر (shār) nor شِہر (shīr) nor شُہر (shūr) nor شْہر (sh hr). The definitions of these two examples are as under:-



(1) شہر: فت مج ش، سک ہ۔ امذ۔ بڑی بستی، نگر، دیہہ / دیہات سے متمیز جہاں کھیتوں کی بجائے بڑا کاروبار، بڑے بازار ہوں۔

Shehr: fatha majhūl shīn, skūn ha, ism mūẕkkar, bṛī bastī, nagar, dayhh/dayhāt say mutmīīz jhāṉ khaytuṉ kī bjāay bṛā kārwbār, bṛay bāzār hoṉ.

"City: Horizontal line above *shīn*, amputation on *ha*, Noun masculine. Municipality, town, opposite to village or villages, where there are businesses instead of farming."



(2) صحرا: فت مج ص، سک ح۔ امذ۔ ریگستان، دشت، ویرانہ، بیابان، جنگل۔

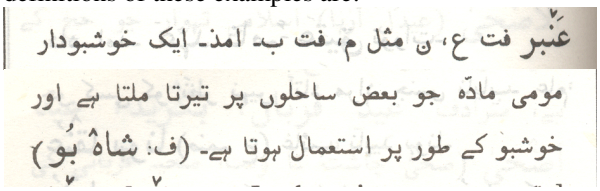
Ṣhrā: fatha majhūl ṣād, skūn ḥa, ism mūẕkkar, raygistān, dsht, wīrānā, bayābān, jangle.

Desert: Horizontal line above *ṣād*, amputation on *ḥa*, Noun masculine. Arid region, sandy area, wasteland, jangle.

The shape of Arabic Zwarkay (U+0659) used in Pashtu is exactly equal to the shape of Fatha Majhool. If agreed by the linguists, it is recommended that the definition of Fatha Majhool for its use in Urdu may be included in U+0659. Alternatively, a new symbol for it can be included in Unicode.

11. Ulta Jazm ◌ٗ

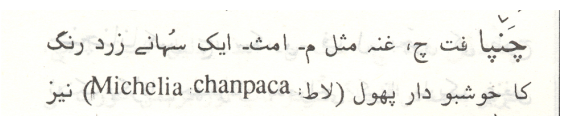
The shape of Ulta Jazm is similar to "Arabic Vowel Sign Inverted Small V Above" (U+065B). It is used when Urdu character Meem (م) appears in the same of Noon (ن). The examples include:- عنبر and چنپا. In both examples, the sound of character ن is of م. The definitions of these examples are:-



(1) عنبر: فت ع، ن مثل م، فت ب۔ امذ۔ ایک خوشبودار مومی مادہ جو بعض ساحلوں پر تیرتا ملتا ہے اور خوشبو کے طور پر استعمال ہوتا ہے۔

'mbar: fatha 'an, nūn misl mīm, fatha ba, ism mūẕkkar. aik khūshbūdār momī mādh jo b'aẕ sāḥilūṉ par tayrtā ha awr khūshbū kay ṭur par ist'amāl hotā ha.

"Ambergris: tiled line above *'an*, *nūn* sound-like *mīm*, tiled line above *ba*, Noun masculine. A fragranced wax-like material which sometimes floats on beaches and used as perfume."



(2) چنپا: فت چ، غنہ مثل م (چَم پا) امث۔ ایک سہانے زرد رنگ کا خوشبو دار پھول۔

Chmpā: fatha cheh, ghunha misl mīm, (cham pā), ism muanus. Ayk suhānay zrd rṇg kā khūshbūdār phūl.

"Jasmine: tiled line above *cheh*, *ghunha* sound-like *mīm*, Noun feminine. A pleasant yellowish coloured fragranced flower."

12. Bibliography

[1] Dr. G. C. Narang, *اردو زبان اور لسانیات*, Sang-e-Meel Publishers, Lahore, 2007

[2] J. T. Platts, *اُردو کلاسیکی ہندی اور انگریزی ڈکشنری*, Urdu Science Board, Lahore, 2005

[3] K. A. Hameed, *جامع اللغات*, Urdu Science Board, Lahore, 2003

[4] M. A. Haq, *اردو لغت کلاں*, Tarraqi-e-Urdu Board (Urdu Dictionary Board), Karachi, 1979-2007

[5] M. A. Haq, *لغتِ کبیر*, Anjuman Taraqqi-e-Urdu Pakistan, Lahore, 1977

[6] M. A. Haq, *قواعدِ اردو*, Lahore Academy, Lahore

[7] M. Noor ul Hassan, *نور اللغات*, National Book Foundation, Islamabad, 1976

[8] M. S. A. Dehlvi, *فرہنگِ آصفیہ*, Sheikh Ghulam Ali and Sons, Lahore, 1918

[9] M. S. T. H. Rizvi, *لغاتِ کشوری*, Sang-e-Meel Publishers, Lahore, 2003

[10] R. H. Khan, *انشا اور تلفظ*, Izhar Sons, Lahore, 1993

[11] R. H. Khan, *اردو املا*, Majlis-e-Tarraqui-e-Adab, Lahore, 2007

[12] R. H. Khan, *املا کیسے لکھیں*, Izhar Sons, Lahore, 2007

[13] R. H. Khan, *عبارت کیسے لکھیں*, Izhar Sons, Lahore, 2007

[14] S. R. Faruqi, *لغاتِ روزمرہ*, City Press Book Shop, Karachi, 2003

[15] S. H. Haqqui, *فرہنگِ تلفظ*, National Language Authority, Islamabad, 2002

[16] S. Badar ul Hassan, *صحتِ املاء*, Dar ul Noor, Lahore, 2005

[17] T. Hashmi, *اصلاحِ تلفظ و املاء*, Al Qamar Enterprises, Lahore

[18] V. Neufeldt, *Webster's New World Dictionary, Third College Edition*, Webster's New World Dictionaries, New York, 1988

13. Acknowledgement

1. Dr. Khurshid Rizvi
2. Mr. Hafiz Safwan Muhammad Chohan
3. Mr. Wasi Ullah Khokhar

14. References

[1] en.wikipedia.org/Urdu

[2] http://www.nla.gov.pk/beta/images/alphabetsfull.gif

[3] V. Neufeldt, *Webster's New World Dictionary, Third College Edition*, Webster's New World Dictionaries, New York, 1988.

[4] A. S. Hornby, *Oxford Advanced Learner's Dictionary of Current English, Fourth Edition*, Oxford University Press, 1989

[5] S. H. Haqqui, *فرہنگِ تلفظ*, National Language Authority, Islamabad, 1996

[6] With proper marks clearly defining the pronunciation or sound

[7] اسم or noun

[8] مذکر or masculine

[9] Though included in Unicode, but, no font still supports it and a few symbols mentioned in the next entries.

[10] فرہنگِ تلفظ and اردو لغت کلاں has used a different symbol of Waw Mar'uf.

[11] Arabic zwarakay is proposed to be renamed as "Fatha Majhool"

[12] The correct placement is a circle above Waw

[13] S. H. Haqqui, *فرہنگِ تلفظ*, National Language Authority, Islamabad, 1996

[14] S. H. Haqqui, *فرہنگِ تلفظ*, National Language Authority, Islamabad, 2002

[15] M. A. Haq, *قواعدِ اردو*, Lahore Academy, Lahore, 2007