ISO/IEC JTC 1/SC 2/WG 2 **N2413-4**

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]

Please fill all the sections A, B and C below.

(Please read Principles and Procedures Document for guidelines and details before filling this form.)

See http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **for latest *Form*.**

See http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for latest *Principles and Procedures* document.**

See http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest roadmaps.**

A. Administrative

1. **Title:** **Proposal to add Marks and Digits in Arabic Code Block (for Urdu)**

2. Requester's name: Urdu and Regional Language Software Development Forum, Ministry of Science and Technology, Government of Pakistan

3. Requester type (Member body/Liaison/Individual contribution): National Body

4. Submission date: 30-Apr-2002

5. Requester's reference (if applicable): _____

6. (Choose one of the following:)
This is a complete proposal: Yes
or,
More information will be provided later: _____

B. Technical - General

1. (Choose one of the following:)
   a. This proposal is for a new script (set of characters): No
   Proposed name of script: _____
   b. The proposal is for addition of character(s) to an existing block: Yes
   Name of the existing block: Arabic Code Block (U0600)

2. Number of characters in proposal: 16

3. Proposed category (see section II, Character Categories): A

4. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000): 2 (includes CC)
   Is a rationale provided for the choice? _____
   If Yes, reference: _____

5. Is a repertoire including character names provided? Yes
   a. If YES, are the names in accordance with the 'character naming guidelines in Annex L of ISO/IEC 10646-1: 2000? Yes
   b. Are the character shapes attached in a legible form suitable for review? Yes

6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? Dr. Sarmad Hussain, Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, B Block Faisal Town, Lahore, Pakistan. (sarmad.hussain@nu.edu.pk)
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
   _____

7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? Yes
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? Yes

[1] Form number: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09)

8. Special encoding issues:

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?

Yes

9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? No

If YES explain _____

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? Yes

If YES, with whom? National Language Authority, General Public (through Newspaper Ad)

If YES, available relevant documents: see http://www.dkuug.dk/jtc1/sc2/wg2/docs/n2413-3.pdf

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Yes

Reference: General public doing Urdu computing or publication

4. The context of use for the proposed characters (type of use; common or rare) Common

Reference: General books in Urdu, details presented later in this document

5. Are the proposed characters in current use by the user community? Yes

If YES, where? Reference: General books in Urdu, details presented later in this document

6. After giving due considerations to the principles in *Principles and Procedures document* (a WG 2 standing document) must the proposed characters be entirely in the BMP? _____

If YES, is a rationale provided? _____

If YES, reference: _____

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? Yes for 10. For others not necessary

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? Yes (for 11 chars)

If YES, is a rationale for its inclusion provided? Yes

If YES, reference: Later in the document

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? No

If YES, is a rationale for its inclusion provided? _____

If YES, reference: _____

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? Yes

If YES, is a rationale for its inclusion provided? Yes

If YES, reference: Later in the document

11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)? _____

If YES, is a rationale for such use provided? _____

If YES, reference: _____

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? _____

If YES, reference: _____

12. Does the proposal contain characters with any special properties such as control function or similar semantics? No

If YES, describe in detail (include attachment if necessary) _____

13. Does the proposal contain any Ideographic compatibility character(s)? No

If YES, is the equivalent corresponding unified ideographic character(s) identified? _____

If YES, reference: _____

1. Proposal

This paper presents 6 marks and 10 digits which are used for Urdu publications and have been included in Urdu Zabta Takhti (UZT) 1.01, the national standard of Government of Pakistan for Urdu [1],[2],[3]. The 6 marks are completely missing in ISO/IEC 10646 standard and the 10 digits are present in ISO/IEC 10646 but need to be re-coded for reasons to be discussed. Most of the other characters in UZT 1.01 are either already in the ISO standard, or have been proposed [4]. There are still a few more characters and marks in UZT 1.01 not proposed for ISO standard earlier and not included in this proposal, as they are still being debated and will be put forth after the debate has been settled.

The sixteen marks being proposed in this proposal are commonly used in Urdu printing. They are semantically mixed, representing punctuation and linguistic meaning.

1. URDU MISRA SIGN

Urdu poetry is normally written as couplets. A couplet is called a ‘shayr’ in Urdu and each line in the couplet is called a ‘Misra’. Figure 1 below shows a ‘ghazal’ (special kind of poem) in Urdu divided into couplets. Each arrow in this figure shows a ‘Misra’ and the two arrows combined represent two ‘Misra’ or one ‘shayr’.

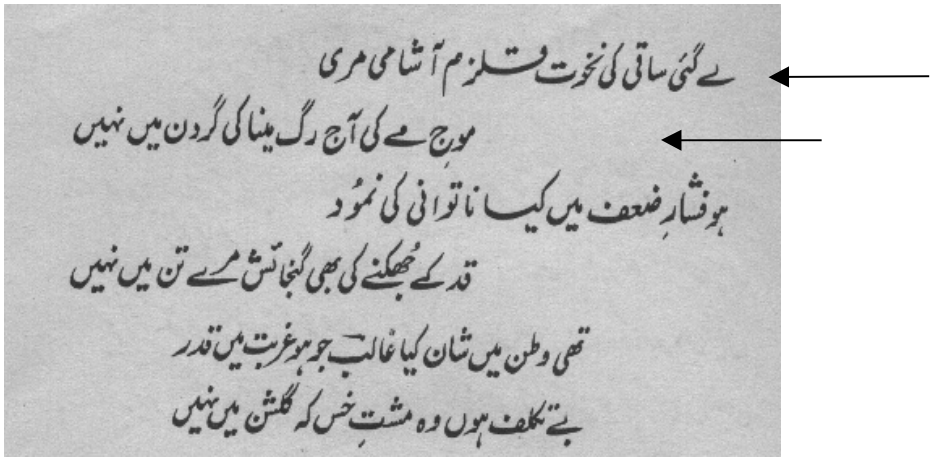

Figure 1 : Couplets (each composed of two 'Misra') in Urdu [5]

When one line of the couplet (i.e. 'misra') has to be quoted within text of Urdu, a special symbol is used to represent it. This symbol is written in a new line, followed by the 'misra', as shown in Figure 2.

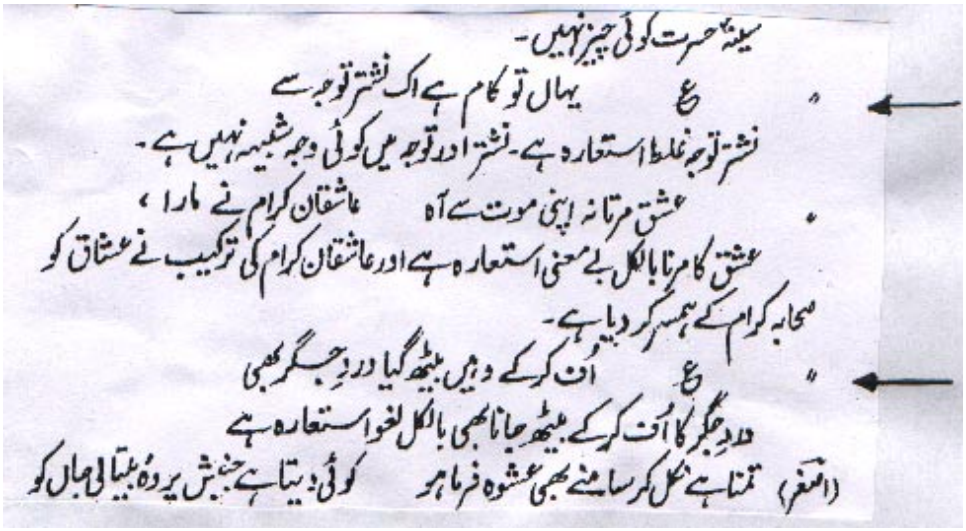

Figure 2: Usage example of 'Misra' sign in Urdu [6] (p. 340)

The glyph is similar in shape to ARABIC LETTER AIN (0639), however, there is a distinct difference in shape (the tail end of AIN curls back into the letter), which is not difference in writing style but actually has semantically different meaning. AIN with the curl may not be used for normal AIN (or vice versa). As both ARABIC LETTER AIN and ARABIC SIGN MISRA have different glyphs and are used concurrently in Urdu, they may not be represented by the same code. Therefore, it is proposed to add this glyph into the ISO standard.

Glyph	Name	Comments	Example
؏	**ARABIC SIGN MISRA**	Used to indicate inclusion of a single line (‘misra’) of a couplet (‘shayr’) of Urdu poetry in prose/text. (UZT 1.01 code : A9)	See Figure 2

2. URDU SAFAH SIGN

When quoting page number in text, footnote or reference ‘p.’ or ‘pp.’ abbreviations are used in English to indicate single or multiple page numbers (e.g. ‘pp. 35-45’ refers to pages 35 till 45). Urdu does not have corresponding abbreviations, but uses a special symbol, the Safah sign, to indicate page number. This sign is formed by the stroke used for the head of ARABIC LETTER SAD (0635), as indicated in Figure 3. below.

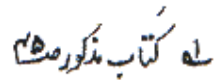

Figure 3: Safah Sign in Urdu to indicate reference to a page [6] (p. 326)

Again, there is a distinct difference in shape, which is not difference in writing style but actually has semantically different meaning. ARABIC LETTER SAD may not be used for ARABIC SIGN SAFAH (or vice versa). As both ARABIC LETTER SAD and ARABIC SIGN SAFAH have different glyphs and are used concurrently in Urdu, they may not be represented by the same code. Therefore, it is proposed to add this glyph into the ISO standard. The stroke of SAFAH sign may be elongated to write page number over it, as indicated in Figure 3.

Glyph	Name	Comments	Example
ﷺ	**ARABIC SIGN SAFAH**	Used to indicate the page number (UZT 1.01 code : AB)	See Figure 3

3. URDU NUQTATAIN

This sign functions similar function to English colon punctuation mark. It is written as two square dots followed by a tapered hyphen-like mark, as shown in Figure 4.

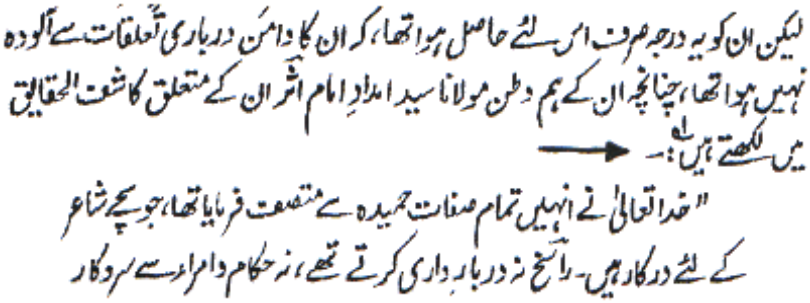

Figure 4 : Urdu Nuqtatain [6] (p. 326)

Nuqtatain may not be represented by combining two different codes (colon + hyphen) for two reasons This mark is semantically perceived as a single unit and is not intuitively two marks for Urdu writers and readers. Secondly, the hypen is not tapered and is required to represent the minus sign in urdu (as in English).

Glyph	Name	Comments	Example
	**ARABIC NUQTATAIN URDU**	Used in a similar fahsion as colon in English (UZT 1.01 code : 3A)	See Figure 4

5. URDU JAZM

In written Urdu, vowels are specified by aerab (ARABIC KASRA, ARABIC FATHA and ARABIC DAMMA, 0650, 064E and 064F respectively) for short vowels and aerab plus letters (limited to ARABIC LETTER ALEF, ARABIC LETTER WAW and ARABIC LETTER FARSI YEH, 0627, 0648 and 06CC respectively) for long vowels. Aerab are not written in normal text of Urdu. Readers are familiar with Urdu and can recreate the vowels just through the consonantal sequence provided. As vowels are not explicitly written most of the times, it is hard to mark syllable boundries (determine onset and coda consonants). Jazm is used to indicate that a consonant occurs in the coda (possibly part of a consonant cluster in coda of the syllable). It is similar, semantically, to ARABIC SUKUUN character in the standard. However, Urdu Jazm has a disctinctly different shape from ARABIC SUKUUN, given in Figure 5.

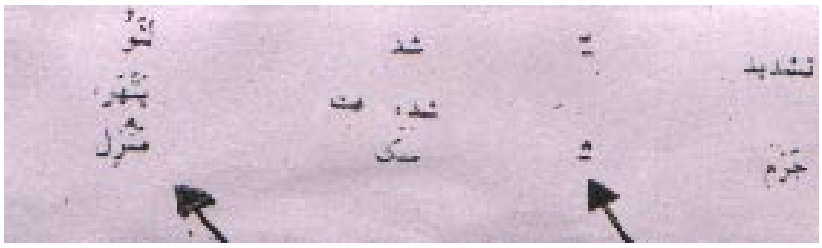

Figure 5 : Urdu Jazm [7]

There are instances, especially in Urdu translations of Quran, where both Arabic and Urdu and written together, e.g. see Figure 6. Though semantically same, if both glyphs have to be put in one document separate codes are needed to be assigned to them.

Figure 6 : Urdu Translation of Quran

Glyph	Name	Comments	Example
	**ARABIC JAZM URDU**	Used to indicate a coda consonant in Urdu syllable (UZT 1.01 code : 4D)	See Figure 5

5. ARABIC SMALL HIGH TAH

When writing Quran in Arabic in South-asia, for Urdu speaking people, a small TAH is used to indicate a break in a phrase, e.g. see the (left) end of the Arabic line in Figure 6 above.

Glyph	Name	Comments	Example
	**ARABIC SMALL HIGH TAH**	Used to indicate phrase break in Arabic for Urdu readers (in South Asia) (UZT 1.01 code : 4D)	See Figure 6

6. BISMILLAH LIGATURE

There are a few ligatures which are very frequently used and are not native to Urdu but have been inherited from Arabic, for religious reasons. In Pakistan, it is required by law to begin all official govenment documents with the Bismillah ligature. However, everybody is clear that it is not Urdu, but Arabic, similar to ligatures already in the standard (e.g. FDFA and FDFB). If a person is typing Urdu, and is unfamiliar with Arabic and has Urdu (not Arabic) keyboard, it will be difficult (if not practically impossible) for the user to type this Arabic ligature. For this reason, and for the reason the two ligatures, referred to above, were introduced in the ISO standard, it is also proposed to include another ligature for Bismillah. This ligature is also part of UZT 1.01, the national standard code page of Government of Pakistan for Urdu.

Glyph	Name	Comments	Example
	**ARABIC LIGATURE BISMILLAH ARRAHMAN ARRAHIM**	Government of Pakistan requires all official documents to start with this ligature (UZT 1.01 code : A2)	

7. URDU DIGITS 0 TO 9

Arabic digits are coded in Unicode in two positions. First, from U 0661 to U 0669 where these are named ARABIC-INDIC DIGITS. Second, as U 06F1 to U 06F9 where these are named as EXTENDED ARABIC-INDIC DIGITS. However at both these positions the glyphs do not completely represent the glyphs used in Urdu. In the first of these positions digits 4, 5, 6 and 7 do not represent Urdu glyphs. In the second set, digits 4, 6 and 7 do not represent Urdu glyphs. In view of this, one solution can be to propose only the missing Urdu glyphs of digits 4, 6 and 7 and then use these in complement with the second set i.e. EXTENDED ARABIC-INDIC DIGITS. However this is a non elegant solution and therefore it is proposed to encode all 10 Urdu glyphs anew in contigious position in the ISO 10646 standard. This has been done and the the Urdu digit glyphs are shown below.

Glyph	Name	Comments	Example
۰	**EXTENDED ARABIC-INDIC DIGIT ZERO (URDU)**	(UZT 1.01 code : 30)	
۱	**EXTENDED ARABIC DIGIT ONE (URDU)**	(UZT 1.01 code : 31)	
۲	**EXTENDED ARABIC DIGIT TWO (URDU)**	(UZT 1.01 code : 32)	
۳	**EXTENDED ARABIC DIGIT THREE (URDU)**	(UZT 1.01 code : 33)	
۴	**EXTENDED ARABIC DIGIT FOUR (URDU)**	(UZT 1.01 code : 34)	
۵ h	**EXTENDED ARABIC DIGIT FIVE (URDU)**	(UZT 1.01 code : 35)	
۶	**EXTENDED ARABIC DIGIT SIX (URDU)**	(UZT 1.01 code : 36)	

۷	**EXTENDED ARABIC DIGIT SEVEN (URDU)**	(UZT 1.01 code : 37)	
۸	**EXTENDED ARABIC DIGIT EIGHT (URDU)**	(UZT 1.01 code : 38)	
۹	**EXTENDED ARABIC DIGIT NINE (URDU)**	(UZT 1.01 code : 39)	

2. Conclusion

In this proposal 16 glyphs have been proposed for inclusion into ISO 10646 (Unicode) standard. These glyphs have already been incorporated into the national standard which has been approved by Government of Pakistan. The inclusion of these glyphs into ISO 10646 will make available the entire character set of Urdu thereby enhancing the portability of the language across different platforms supporting ISO 10646. Further these enhancements will facilitate publishing work in Urdu to be carried out in Unicode compatible text processing systems, which is not currently not feasible.

3. References

1. Urdu Computing Standards: UZT 1.01, in *Proceedings of the IEEE International Multi-Topic Conference*, Lahore. S. Hussain and M. Afzal (2001).
2. Urdu Computing Standards: Development of UZT 1.01, in *Proceedings of the IEEE International Multi-Topic Conference*, Lahore. M. Afzal and S. Hussain (2001).
3. Towards Unicode Standard for Urdu, in *Proceedings of 4th Symposium on Multilingual Information Processing (MLIT4)*, Yangon, Myanmar (CICC Japan). K. Zia (1999).
4. Proposal to add Arabic-script honorifics and other Marks (L2/01-425), by Jonathan Kew (SIL International : jonathan_Kew@sil.org).
5. *Diwan-e-Ghalib.* Sang-e-Meel Publications, Lahore, Pakistan (1995), p. 63.
6. *Shayr-ul-Hind.* Maulana Abdus Salaam Nidvi. Ishrat Publishing House, Lahore. (1965)
7. Farhang-e-Talaffuz. Shan-ul-Haq Haqqi. National Language Authority, Islamabad, Pakistan.