ISO/IEC JTC/SC2/WG2

Universal Multiple—Octet Coded Character Set
(UCS)

ISO/IEC JTC/SC2/WG2 N3919
2010.9.15

TITLE: Proposal to Encode Special Scripts and Characters in UCS for Uighur language
SOURCE: China
STATUS: NATIONAL BODY POSITION
ACTION: CONSIDERATION BY WG2 Preliminary Proposal, for collecting comments

----

In order to deal with the incompleteness of information exchanges for Uighur languages , it is quite necessary to supplement **Eight** special scripts of Uighur letters in ISO/IEC 10646 /Unicode (table1-2, below).

Problame 1: **1)** We heve emphasize the importance of the nominal form of Uighur character "ئا" is a indivisible symbol, and that is do not represented by combining symbol "ئ+ا = ئا ", it does not consist with the writing regulations of Uighur letters. We emphasize again it is a completely single and a indivisible symbol . At the same time, the shape Uighur letter "ئا" is different to Arabic letter "ا", thus it has to be separately encoded. Similarly, Uighur character "ئە", "ئى", "ئو" , "ئۆ" , "ئۇ"and "ئۈ" are indivisible symbols and cannot be represented by combining symbol with "ئ", or not dual-joining with U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE. Other one You can see, they are different to Arabic letters "ە", "ى", "و" , "ۆ" , "ۇ" and"ۈ" in shapes and quantity. So These characters represented by sequence of two existing characters is a completely wrong action, such as NFS.

"05F7" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+0627 ARABIC LETTER ALEF

"05F8" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+06D5 ARABIC LETTER AE

"05FB" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+0648 ARABIC LETTER WAW

"05FC" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+06C7 ARABIC LETTER U

"05FD" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+06C6 ARABIC LETTER OE

"05FA" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+0649 ARABIC LETTER ALEF MAKSURA

Therefore, all of these letters need to be encoded separately.

2) The presentation forms of eight nominal forms for Uighur, languages are already encoded UCS (see FBEA -- FBFD in Table 133), but their nominal forms are not encoded till now.

Problame 2: There exist numerous nominal forms in Uighur, Kazakh , Kirghiz and Arabic information processing and exchange, which cause ambiguity, since their presentation forms are different in quantity or shapes.

1) Table 1-1 lists Uighur, Kazakh, Kirghiz and Arabic letters for nominal form in UCS and nominal form in original shape, which are completely different in shape and variant quantity. It certainly produces double ambiguous codes in UCS. Under serial number 2 in table 1-1, nominal form "ﯖ" will be supplemented to Uighur language, because it has four variants: isolated form ("ﯖ " and " ﻩ "), final form ("ﻪ" and "ﯕ"). While the nominal form of Kazakh and Kirghiz letter"ﻩ"(06d5) has two variants: isolated form "ﻩ" and final form "ﻪ ". They are different in shape and variant quantity. If Nominal form of Kazakh and Kirghiz letter "ﻩ"(06d5) replaces Uighur Nominal form"ﯖ   " or isolated form " ﻩ ", there appears ambiguity in information exchange as a result.

2) Likewise the nominal form of Arabic, Kazakh, Kirghiz language under serial number 1,2,3,4,5,6 of table 1-1 are different in shape comparing to nominal forms of Uighur language which have variant quantity and part of member in shapes, therefore **Eight** nominal forms for Uighur language need to be supplemented. For example, under serial number 3, Arabic letter YEH "ﻯ" (U+0649) has two variants: ("ﻯ" and "ﻰ"), but its corresponding character in Kazakh and Kirghiz has four variants: ("ﻯ","ﻰ ","ﻲ" and "ﺋ"). In Uighur, it has **Eight** variants. So Arabic letter U+0649 cannot be used to represent Uighur, Kazakh and Kirghiz letter mentioned above. Therefore, Uighur letter "i" "ﺋﻰ".should be separately encoded suggestion code points to U+05FA.

In addition, U+0648 "ﻭ" is Arabic, Kazakh and Kirghiz letter, which has two variants ("ﻭ" and "ﻮ"), but it cannot be used to represent letter "ﺋﻮ" in Uighur language which has four variants :("ﻭ", "ﻮ", "ﺋﻮ" and "ﯮ"). Although the nominal form of this letter is of Arabic, Kazakh, Kirghiz and Uighur languages, it has different variant quantity and shapes. So, it is necessary to encode Uighur letter "ﺋﻮ"separately, and we suggest to encode it at U+05FB. Some other letters such as U+06c7 "ﺋﯘ", U+06c6 "ﺋﯚ" and U+0627 "ﺋﺎ" have the same problem mentioned above. Therefore all of these letters should be encoded separately.

The Suggested code points for nominal forms of Uighur languages have to be in 05XX 05f7-05ff (may be arrange in other code points) .

Table 1-1: Some Uighur, Kazakh ,Kirghiz and Arabic Letters that produce to Double Ambiguous Codes in 10646

Serial number	Code	Language	Nominal form in 10646	Nominal form –in original	Pronunciation	variants							
						Isolated form	Final form	Initial form	Medial form	Isolated form	Final form	Initial form	Medial form
1	0627	Arabic Kazakh Kirghiz	ا	ا	[a]	ا	ﺎ						
		Uighur	ا	ئا	[a]	ا	ﺎ			ئا	ﺋﺎ		
2	06D5	Kazakh Kirghiz	ە	ە	[e]	ە	ﻪ						
		Uighur	ە	ئە	[æ]	ە	ﻪ			ئە	ﺋﻪ		
3	0649	Arabic	ى	ى		ى	ﻰ						
		Kazakh Kirghiz	ى	ى	[ ]	ى	ﻰ	ﺑ	ﺒ				
		Uighur	ى	ئى	[i]	ى	ﻰ	ﺑ	ﺒ	ئى	ﺋﻰ	ﺋﺪ	ﺌﻤ
4	0648	Arabic Kazakh Kirghiz	و	و	[w. o]	و	ﻮ						
		Uighur	و	ئو	[o]	و	ﻮ			ئو	ﺋﻮ		
5	06C7	Kazakh Kirghiz	ۇ	ۇ	[u]	ۇ	ﯗ						
		Uighur	ۇ	ئۇ	[u]	ۇ	ﯗ			ئۇ	ﺋﯗ		
6	06C6	Kazakh	ۆ	ۆ	[v]	ۆ	ﯙ						
		Uighur	ۆ	ئۆ	[θ]	ۆ	ﯙ			ئۆ	ﺋﯚ		

Table 1-2: Special Scripts to Be Supplemented in ISO/IEC 10646 /Unicode for Uighur languages

Language	Final form	Medial form	Initial form	Isolated form	Nominal form	No
Uighur	ا FE8E ‏ ئا FBEB			ا FE8D ‏ ئا FBEA	ئا 05F7 New	1
Uighur	ه FEEA ‏ ئە FBED			ە FBEC ‏ ە FEE9	ئە 05F8 New	2
Uighur	ئې FBE5 ‏ ې FBF7	FBD1 / FBE7	FBE6 ‏ FBF8	FBE4 ‏ FBF6	ئې 05F9 New	3
Uighur	ئى FEF0 ‏ ى FBFA	FBD2 / FBE9	FBE8 ‏ FBFB	FEEF ‏ FBF9	ئى 05FA New	4
Uighur	ئو FEEE ‏ و FBEF			ئو FEED ‏ و FBEE	05FB New	5
Uighur	FBD8 ‏ FBF1			FBD7 ‏ FBF0	05FC New	6
Uighur	FBDA ‏ FBF3			FBD9 ‏ FBF2	05FD New	7
Uighur	FBDC ‏ FBF5			FBDB ‏ FBF4	05FE New	8

The Suggested code points for **Special Scripts** of Uighur languages have to be in 05XX 05f7-05ff (may be arrange in other code points) .

	059	05A	05B	05C	05D	05E	05F
0	▨	◌ͦ 05A0	◌ 05B0	| 05C0	א 05D0	נ 05E0	וו 05F0
1	◌ 0591	◌ 05A1	◌ 05B1	◌ 05C1	ב 05D1	ס 05E1	וי 05F1
2	◌ 0592	◌ 05A2	◌ 05B2	◌ 05C2	ג 05D2	ע 05E2	יי 05F2
3	◌ 0593	◌ 05A3	◌ 05B3	◌ 05C3	ד 05D3	ף 05E3	׳ 05F3
4	◌ 0594	◌ 05A4	◌ 05B4	◌ 05C4	ה 05D4	פ 05E4	״ 05F4
5	◌ 05C5	◌ 05A5	◌ 05B5	◌ 05C5	ו 05D5	ץ 05E5	▨
6	◌ 0596	◌ 05A6	◌ 05B6	׆ 05C6	ז 05D6	צ 05E6	▨
7	◌ 0597	◌ 05A7	◌ 05B7	◌ 05C7	ח 05D7	ק 05E7	ﭷ **05F7** New
8	◌ 0598	◌ 05A8	◌ 05B8	▨	ט 05D8	ר 05E8	ﭸ **05F8** New
9	◌ 0599	◌ 05A9	◌ 05B9	▨	י 05D9	ש 05E9	ﭹ **05F9** New
A	◌ 059A	◌ 05AA	◌ 05BA	▨	ך 05DA	ת 05EA	ﭺ **05FA** New
B	◌ 059B	◌ 05AB	◌ 05BB	▨	כ 05DB	▨	ﭻ **05FB** New
C	◌ 059C	◌ 05AC	◌ 05BC	▨	ל 05DC	▨	ﭼ **05FC** New
D	◌ 059D	◌ 05AD	◌ 05BD	▨	ם 05DD	▨	ﭽ **05FD** New
E	◌ 059E	◌ 05AE	־ 05BE	▨	מ 05DE	▨	ﭾ **05FE** New
F	◌ 059F	◌ 05AF	◌ 05BF	▨	ן 05DF	▨	▨

	FB5	FB6	FB7	FB8	FB9	FBA	FBB	FBC	FBD	FBE	FBF
0	أَ FB50	ۣ FB60	ۊ FB70	چ FB80	ک FB90	ݐ FBA0	ﮀ FBB0			و FBE0	ئۇ FBF0
1	أَ FB51	ۣ FB61	ۊ FB71	چ FB81	ک FB91	ݑ FBA1	ﮁ FBB1			و FBE1	ئۇ FBF1
2	پ FB52	ٿ FB62	ج FB72	ڊ FB82	گ FB92	ݒ FBA2				ۇ FBE2	ئۆ FBF2
3	پ FB53	ٿ FB63	ج FB73	ڊ FB83	گ FB93	ݓ FBA3			ݟ FBD3	ۇ FBE3	ئۆ FBF3
4	ﭔ FB54	ٺ FB64	چ FB74	ڌ FB84	گ FB94	ﺔ FBA4			ﯔ FBD4	ﯼ FBE4	ئۆ FBF4
5	ﭕ FB55	ٺ FB65	چ FB75	ڌ FB85	ک FB95	ﺔ FBA5			ﯕ FBD5	ﯽ FBE5	ئۆ FBF5
6	پ FB56	ڻ FB66	ج FB76	ڎ FB86	گ FB96	ه FBA6			ﯖ FBD6	ﯦ FBE6	� ئې FBF6
7	پ FB57	ڻ FB67	چ FB77	ڎ FB87	گ FB97	ﻫ FBA7			ۇ FBD7	ﯧ FBE7	ﮗ ئې FBF7
8	ﭘ FB58	ڻ FB68	چ FB78	ڈ FB88	ﮗ FB98	ﺍ FBA8			ۇ FBD8	ﯨ FBE8	ﯘ FBF8
9	ﭙ FB59	ڻ FB69	چ FB79	ڈ FB89	ﮗ FB99	ﻬ FBA9			و FBD9	ﯩ FBE9	ئى FBF9
A	ﭚ FB5A	ڤ FB6A	ﭺ FB7A	ڑ FB8A	ﮒ FB9A	ه FBAA			و FBDA	ﯪ FBEA	ئى FBFA
B	ﭛ FB5B	ﭢ FB6B	چ FB7B	ڑ FB8B	ﮓ FB9B	ه FBAB			و FBDB	ﯫ FBEB	ﻧ FBFB
C	ﭜ FB5C	ﭣ FB6C	چ FB7C	ﮌ FB8C	ﮔ FB9C	ﻫ FBAC			ﯜ FBDC	ﯬ FBEC	ی FBFC
D	ﭝ FB5D	ﭤ FB6D	چ FB7D	ﮍ FB8D	ﮕ FB9D	ﻬ FBAD			ﯝ FBDD	ﯭ FBED	ی FBFD
E	ﭞ FB5E	ﭥ FB6E	چ FB7E	ک FB8E	ں FB9E	ﮰ FBAE			ﯞ FBDE	ﯮ FBEE	ﯟ FBFE
F	ﭟ FB5F	ﭦ FB6F	چ FB7F	ک FB8F	ں FB9F	ﮯ FBAF			ﯟ FBDF	ﯯ FBEF	ﯟ FBFF

Names and suggested code points for the eight letters to be supplemented:

ئا 05F7   Arabic   letter   ALEF   for   Uighur

ئە   05F8   Arabic   letter   AE   for   Uighur

ئى   05F9   Arabic   letter   YEH   for   Uighur

ئې 05FA   Arabic   letter   YEH   for   Uighur

ئو 05FB   Arabic   letter   WAW   for   Uighur

ئۇ 05FC   Arabic   letter   U   for   Uighur

ئۆ 05FD   Arabic   letter   OE   for   Uighur

ئۈ 05FE   Arabic   letter   OU   for   Uighur

Uighur alphabet for children (nominal forms of Uighur letters):

Proposed eight special letters in Uighur words:

ئا : ئالىم , ئاستا، ئادالەت ، ئىئانە ، ئاسراق

ئە : ئەمگەك، ئەقىل ، ئەسئەت، ئەدەب

ئې : ئېبىق ، سېسسق ، دېتال، ئېرلان ، ئېرا

ئى : ئىشىك، ئىلمى، رۇقئى، رۇئايەت، كۆزى

ئو : ئوغاق، روشەن ، ئوغلاق، ئويغاق، ئويلان

ئۇ : ئۇيغۇر، ئۇچۇر، ئۇزۇن ، مە سئۇل، رۇسۇل

ئۆ : ئۆدەك، كۆل، گۆرۆ، ئۆي

ئۈ : ئۈزۈم، ئۈرۈمچى، كۈز، ئۈششۈك

Below are the pictures from Uighur alphabet for children (eight letters which should be supplemented):

ISO/IEC JTC 1/SC 2/WG 2

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 (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html

for guidelines and details before filling this form.

Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.

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

A. Administrative

1. **Title:**	Proposal to Encode Special Scripts and Characters in UCS for Uighur language
2. Requester's name:	China
3. Requester type (Member body/Liaison/Individual contribution):	Member body
4. Submission date:	2010-09-21
5. Requester's reference (if applicable):	No
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: Uighur

    b. The proposal is for addition of character(s) to an existing block: *Yes*

        Name of the existing block: *Uighur*

2. Number of characters in proposal: *8*

3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	X B.1-Specialized (small collection)		B.2-Specialized (large collection)	
C-Major extinct	D-Attested extinct		E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic		G-Obscure or questionable usage symbols	

4. 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 P&P document?   *Yes*

    b. Are the character shapes attached in a legible form suitable for review?   *Yes*

5. Fonts related:

    a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?

    Jpg files were provided by the Xiniiang University, China. Font will be provided later.

    b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):

        chenzh@cesi.ac.cn   wushour@xju,edu.cn

6. 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

7. 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)?	No

8. 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/UCD.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?	Yes
If YES explain WG2n2820, WG2#45 WG2n3819,WG2#56	

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? People in Xinjiang Autonomous Region, China.	
If YES, available relevant documents:	

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	No
Reference:	

4. The context of use for the proposed characters (type of use; common or rare)	common
Reference: For example in education: Uighur alphabet for children.	

5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference: For example, Xinjiang Uighur Autonomous Region, China.	

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	No
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

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	Yes
If YES, is a rationale for its inclusion provided?	Yes
If YES, reference: See problem 1 in the proposal	

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?   *No*

       If YES, is a rationale for its inclusion provided?

          If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences?   *No*

    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: