ISO/IEC JTC/SC2/WG2

Universal Multiple—Octet Coded Character Set
(UCS)

ISO/IEC JTC/SC2/WG2N3819
2010.04.15

TITLE: Preliminary Proposal for Encoding Special Scripts and Characters in UCS for Uighur, Kazakh and Kirgiz
SOURCE: China
STATUS: NATIONAL BODY POSITION
ACTION: Consideration by WG2, for collecting comments

In order to deal with the incompleteness of information exchanges for Uighur, Kazakh, Kirghiz and Arabic languages and occurrence of double ambiguous code, it is quite necessary to supplement Uighur, Kazakh, Kirghiz letters in table1-2.

We understand, the nominal forms with all the presentation forms, which other one different in both of their variant quantity and shape are not entirely identical, giving rise to double ambiguous code. In order to deal with ambiguity caused by letters in different language texts, which have the same shape in nominal forms, but have different quantities or shapes in deformed presentation forms, supplementation is necessary. There are some examples about it in UCS. Such as:

“ ى ” (U+0649) in table 15 and “ ى ” (U+06CC) in table 16;
“ف” (U+0641) in table 15 and “ ڧ ” (U+06A7) in table 16;
“ ھ ” (U+0647) in table 15 and “ ھ ” (U+06BE) in table 16;
“ ع ” (U+0621) in table 15 and “ ۼ ” (U+0674) in table 16;
“ أ ” (U+0623) in table 15 and “ أ ” (U+0622) in table 16;
“ إ ” (U+0625)in table15 and “ إ ” (U+0673) in table 16;

The Nominal form shapes of these letters in corresponding pairs are same, but quantity of presentation forms or shapes are different. Thus, different nominal forms are assigned code points respectively for all of them. At present, there exist numerous nominal forms in Uighur, Kazakh and Kirghiz 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 " ﮒ " will be supplemented to Uighur language, because it has four variants: isolated form (" ﮒ " and " ﮓ "), final form (" ﮔ " and " ﮕ"). While the nominal form of Kazakh and Kirghiz letter" ﯓ "(06d5) has two variants: isolated form " ﯓ " and final form " ﯔ". They are different in shape and variant quantity. If Nominal form of Kazakh and Kirghiz letter " ﯓ "(06d5) replaces Uighur Nominal form" ﮒ " or isolated form " ﮓ ", 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 five 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","ﯼ" and " ﯽ "). 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, Kazakh and Kirghiz letter "i" should be separately encoded.
We suggest to encode a character in U+076E for Kazakh and Kirghiz letter"ﯼ", and a character in U+0772 for Uighur "ﯾ". In addition, U+0648 " ﻮ " is Arabic, Kazakh and Kirghiz letter, which has two variants (" ﻮ " and " ﯙ"), but it cannot be used to represent letter "ۇ" in Uighur language which has four variants :(" ﻮ ", " ﯙ", "ۇ" and "ﯘ"). 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 "ۇ"separately, and we suggest to encode it at U+076F. Some other letters such as U+06c7 " ﯗ", U+06c6 " ﯖ" and U+0627 " ﺋﺎ" have the same problem mentioned above. Therefore all of these letters should be encoded separately.

3) We here emphasize the importance of the nominal form of Uighur letter " ﺋﺎ" which is an indivisible symbol and cannot be represented by combining symbol " ﺋ + ﺎ = ﺋﺎ", it does not consist with the writing regulations of Uighur letters. Here we can see that the shape Uighur letter " ﺋﺎ" is different to Arabic letter " ﺍ ", thus it has to be separately encoded. Similarly, Uighur letters " ﮒ ", "ﯨ", "ﯘ", " ﯗ" and " ﯖ" are indivisible symbols and cannot be represented by combining symbol with " ﯩ ". You can see, they are different to Arabic letters " ﻩ ", "ى", " ﻮ ", " ﯞ " and " ﻕ " in shapes and quantity. Therefore, all of these letters need to be encoded separately otherwise they will cause double ambiguities.

4) The presentation forms of eight nominal forms for Uighur, Kazakh and Kirghiz languages are already encoded UCS (see FBEA -- FBFD in Table 133), but their nominal forms are not encoded.

Since 06XX and 07XX are already occupied or reserved for Arabic characters lately,

the supplemented code points for nominal forms of Uighur, Kazakh and Kirghiz languages have to be placed in 07XX in 076B-0772.

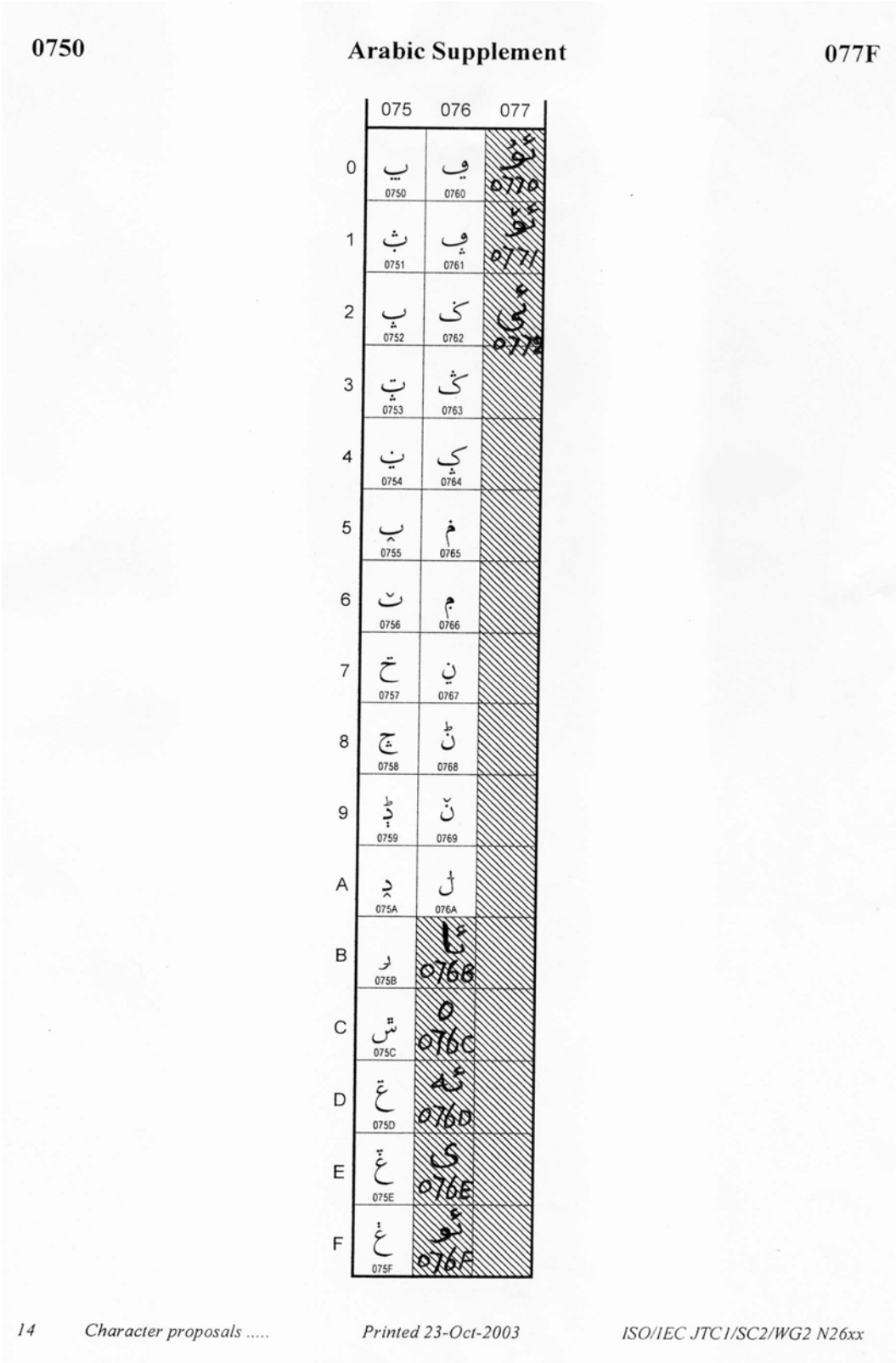
Table 1-1: Uighur, Kazakh and Kirghiz Letters that produce to Double Ambiguous Codes in 10646

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	ـئا		
2	06D5	Kazakh Kirghiz	ە	ە	[e]	ە	ە						
		Uighur	ە	ئە	[æ]	ە	ە			ئە	ـئە		
3	0649	Arabic	ى	ى		ى	ى						
		Kazakh Kirghiz	ى	ى	[ ]	ى	ى	ـىـ	ـىـ				
		Uighur	ى	ئى	[i]	ى	ى	ـىـ	ـىـ	ئى	ـئى	ئـى	ـئـى
4	0648	Arabic Kazakh Kirghiz	و	و	[w.o]	و	و						
		Uighur	و	ئو	[o]	و	و			ئو	ـئو		
5	06C7	Kazakh Kirghiz	ۇ	ۇ	[u]	ۇ	ۇ						
		Uighur	ۇ	ئۇ	[u]	ۇ	ۇ			ئۇ	ـئۇ		
6	06C6	Kazakh	ۆ	ۆ	[v]	ۆ	ۆ						
		Uighur	ۆ	ئۆ	[θ]	ۆ	ۆ			ئۆ	ـئۆ		

Table 1-2: Uighur, Kazakh and Kirghiz Letters to Be Supplemented

Language	Final form	Medial form	Initial form	Isolated form	Nominal form	No
Uighur	ﺎ ﺋﺎ FE8E FBEB			ﺍ ﺋﺎ FE8D FBEA	ﺋﺎ 076B New	1
Kazakh kirghiz	ﻫ FEEA			ە FEE9	ە 076C New	2
Uighur	ﻫ ﺌﻪ FEEA FBED			ﺋﻪ ە FBEC FEE9	ﺋﻪ 076D New	3
Kazakh kirghiz	ﻯ FEF0	ـﻧـ FBE9	ﺩ FBE8	ﻯ FEEF	ﻯ 076E New	4
Uighur	ﻯ ﺋﻰ FEF0 FBFA	ـﺌـ FBD2 ـﻧـ FBE9	ﺩ ﺋﻨ FBE8 FBFB	ﻯ ﺋﻰ FEEF FBF9	ﺋﻰ 0772 New	5
Uighur	ﻭ ﺋﻮ FEEE FBEF			ﻭ ﺋﻮ FEED FBEE	ﺋﻮ 076F New	6
Uighur	ﯗ ﺋﯘ FBD8 FBF1			ﯗ ﺋﯘ FBD7 FBF0	ﺋﯘ 0770 New	7
Uighur	ﯙ ﺋﯙ FBDA FBF3			ﯙ ﺋﯙ FBD9 FBF2	ﺋﯙ 0771 New	8

Suggested code points for letters which should be supplemented.



FB50 **Arabic Presentation Forms-A**

	FB5	FB6	FB7	FB8	FB9	FBA	FBB	FBC	FBD	FBE	FBF
0	ﭐ FB50	ﭰ FB60	ﮀ FB70	ﮠ FB80	ﮐ FB90	ﮰ FBA0	ﯓ FBB0			ﯠ FBE0	ﯰ FBF0
1	ﭑ FB51	ﭱ FB61	ﮁ FB71	ﮡ FB81	ﮑ FB91	ﮱ FBA1	ﯔ FBB1			ﯡ FBE1	ﯱ FBF1
2	ﭒ FB52	ﭲ FB62	ﮂ FB72	ﮢ FB82	ﮒ FB92	﮲ FBA2				ﯢ FBE2	ﯲ FBF2
3	ﭓ FB53	ﭳ FB63	ﮃ FB73	ﮣ FB83	ﮓ FB93	﮳ FBA3			ﴓ FBD3	ﯣ FBE3	ﯳ FBF3
4	ﭔ FB54	ﭴ FB64	ﮄ FB74	ﮤ FB84	ﮔ FB94	﮴ FBA4			ﴔ FBD4	ﯤ FBE4	ﯴ FBF4
5	ﭕ FB55	ﭵ FB65	ﮅ FB75	ﮥ FB85	ﮕ FB95	﮵ FBA5			ﴕ FBD5	ﯥ FBE5	ﯵ FBF5
6	ﭖ FB56	ﭶ FB66	ﮆ FB76	ﮦ FB86	ﮖ FB96	﮶ FBA6			ﴖ FBD6	ﯦ FBE6	ﯶ FBF6
7	ﭗ FB57	ﭷ FB67	ﮇ FB77	ﮧ FB87	ﮗ FB97	﮷ FBA7			ﴗ FBD7	ﯧ FBE7	ﯷ FBF7
8	ﭘ FB58	ﭸ FB68	ﮈ FB78	ﮨ FB88	ﮘ FB98	﮸ FBA8			ﴘ FBD8	ﯨ FBE8	ﯸ FBF8
9	ﭙ FB59	ﭹ FB69	ﮉ FB79	ﮩ FB89	ﮙ FB99	﮹ FBA9			ﴙ FBD9	ﯩ FBE9	ﯹ FBF9
A	ﭚ FB5A	ﭺ FB6A	ﮊ FB7A	ﮪ FB8A	ﮚ FB9A	﮺ FBAA			ﴚ FBDA	ﯪ FBEA	ﯺ FBFA
B	ﭛ FB5B	ﭻ FB6B	ﮋ FB7B	ﮫ FB8B	ﮛ FB9B	﮻ FBAB			ﴛ FBDB	ﯫ FBEB	ﯻ FBFB
C	ﭜ FB5C	ﭼ FB6C	ﮌ FB7C	ﮬ FB8C	ﮜ FB9C	﮼ FBAC			ﴜ FBDC	ﯬ FBEC	ﯼ FBFC
D	ﭝ FB5D	ﭽ FB6D	ﮍ FB7D	ﮭ FB8D	ﮝ FB9D	﮽ FBAD			ﴝ FBDD	ﯭ FBED	ﯽ FBFD
E	ﭞ FB5E	ﭾ FB6E	ﮎ FB7E	ﮮ FB8E	ﮞ FB9E	﮾ FBAE			ﴞ FBDE	ﯮ FBEE	ﯾ FBFE
F	ﭟ FB5F	ﭿ FB6F	ﮏ FB7F	ﮯ FB8F	ﮟ FB9F	﮿ FBAF			ﴟ FBDF	ﯯ FBEF	ﯿ FBFF

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

ئا	076B	Arabic	letter	ALEF	for	Uighur
ە	076C	Arabic	letter	AE	for	Kazakh,Kirghiz
ئە	076D	Arabic	letter	AE	for	Uighur
ى	076E	Arabic	letter	YEH	for	Kazakh,Kirghiz
ئى	0772	Arabic	letter	YEH	for	Uighur
ئو	076F	Arabic	letter	WAW	for	Uighur
ئۇ	0770	Arabic	letter	U	for	Uighur
ئۆ	0771	Arabic	letter	OE	for	Uighur

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

Proposed eight special letters in Uighur words:

ئا : ئالىم ,ئاستا،ئادالەت ،ئىمانە،ئامراق

ئە: ئەمگەك،ئەقىل ،ئەسئەت،ئەدەب

ئې: ئېيىق،سېسىق،دېتال،ئېرلان،ئېرا

ئى: ئىشلىك،ئىلمى،رۇقئى،رىۋايەت،كۆزى

ئو: ئوغاق،روشەن،ئوغىلاق،ئويغاق،ئويلان

ئۇ: ئۇيغۇر،ئۇچۇر،ئۇزۇن،مەسئۇل،رۇسۇل

ئۆ: ئۆدەك،كۆل،گۈرۈچ،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 for Encode Special Scripts and Characters in UCS for Uighur, Kazakh and Kirgiz
2. Requester's name: China
3. Requester type (Member body/Liaison/Individual contribution): National Body
4. Submission date: 2010.4.15
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: No

B. Technical – General

1. Choose one of the following:
   - a. This proposal is for a new script (set of characters): *Yes*
     - Proposed name of script: Uighur, Kazakh and Kirgiz
   - b. The proposal is for addition of character(s) to an existing block: *No*
     - Name of the existing block:
2. Number of characters in proposal: *8*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   - A-Contemporary: Z
   - 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. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Jpg files were provided by the Xiniiang University, China. Font will be provided later.*

   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *wushour@xju.edu.cn, chenzh@cesi.ac.cn, chenzh-zhuang@163.com*
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.

[1] Form number: N3152-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *Yes*
   If YES explain *WG2n2820, WG2#45*
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? *Yes*
   Reference: *For example,Table2 Orkhun-Yenisey Alphabet,Table-3 Old Turk(Orkhun)Alphabet,Table-4:"Iski Turk Yazitlari",Yusayin Namiq Orqun,Yurkiye,Istanbul*
4. The context of use for the proposed characters (type of use; common or rare) *People through Interner*
   Reference: *For example ,Uyghur alphabet.*
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 table1-1,1-2,FB50 Axabic Presentation from S-A*
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: