L2/13-071

ISO/IEC JTC1/SC2/WG2N
2013-3-27

Proposal to Encode the Uyghur Script in ISO/IEC 10646

Omarjan Osman
Nagaoka University of Technology
Kamitomioka 1603-1, Nagaoka Shi, Niigata 940-2188, Japan.
Global Information Infrastructure Laboratory.
s075386@stn.nagaokaut.ac.jp.

March 27, 2013

Abstract

Uyghur script, which is based on Aramaic alphabet is composed of phonetic characters. The Uyghur writing system that had been used in the Turkistan area in Central Asia since around the eighth century until the end of the nineteenth century is completely forgotten in modern days. Up to now, it has not become the object of information processing. However, this writing system is a direct ancestor of many writing systems of the East Asian cultures like Mongolian and Manchurian. Moreover, it has been used as a medium to record a lot of historical documents that have high cultural value. Authors want to contribute to the creation of technology as the basis for the preservation and utilization of the historical Uyghur documents by establishing a character code for the writing system. This article introduces the results of our study on this issue and proposes a Uyghur character code design together with a glyph table design, and some background ideas behind these designs. The Uyghur character codes are not yet included in International standard ISO / IEC 10646 and Unicode. In this proposal, the authors propose a design of Uyghur character code and glyph table.

Keyword: Uyghur character, character code, ISO/IEC 10646, glyph and font

1 Introduction

This is a proposal to encode the Uyghur script in Roadmap to the Supplementary Multilingual Plane (Plane 1, [7]) of the Universal Character Set (ISO/IEC 10646).

1.1 Background and purpose of proposal

Documents written in Uyghur characters are archived in museums and universities of many countries. However, the deterioration of them is unavoidable no matter how well they are stored. Therefore, they are often archived as image files by scanning[1], but it is impossible to handle the documents as text, and thus quantitative analysis of their contents is restricted. When we handle historical documents and materials, however, subtle differences related to fluctuations of usage and/or description may sometimes be the key to determination of many situations and, therefore, experts have carried out such quantitative analyses of the Uyghur documents[2],[3]. The result of those analyses is the fruit gained through the investigation and interpretation of those documents and materials that have been made by the experts for a long time, and thus it goes without saying that such an analysis method is indispensable in the future as well. If the character code is

Table 1: The ISO 639 Language Code of the Selected Language.

Language name	ISO 639-1	ISO 639-2
Aramaic	---	*arc*
Sogdian (Sogdish)	---	*sog*
Uyghur	---	*uig*
Arabic	ar	*ara*
Mongolian	mn	*mon*
Manchu	---	*mnc*
Buryat	---	*bua*
Todo(Oirat)	---	—

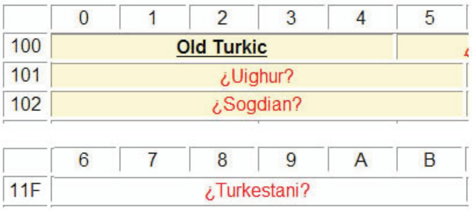
Table 2: Various central Asian scripts in international standards.

Script Name	ISO 15924	ISO/IEC10646
Orkhon	Orkh	*U+10C00∼U+10C4F*
Manichaean	Mani	*There is a formal proposal*[23]
Uyghur	Unregistration	*There is no formal proposal*[24]
Sogdian	Unregistration	*There is no formal proposal*
Mongolian	(Mong)	*U+1800∼U+1842*
Manchu(Sibe)	Unregistration	*U+185D∼U+1877*
Todo(Oirat)	Unregistration	*U+1843∼U+185C*

established, it is possible to compile the texts as digital archives in a compatible form and conduct quantitative analysis for detecting such subtle fluctuations easily and accurately. This issue is mentioned in Section 4 of this article. In addition, the digitalization would make it possible to substitute easily the Modern Uyghur script and Latin script, which could contribute to broadening the perspective of researchers and users.

For those reasons, many characters of historic scripts, such as sacred Egyptian characters, cuneiform characters, Linear A, and Faistos disc characters, were proposed and registered in the Supplementary Multilingual Plane (SMP) of ISO/IEC 10646 (UCS; Universal Coded Character Set), which is the international standard for character codes[4]. ISO/IEC 10646 adopts 32-bit-based character encoding that has been developed for the purpose of compiling characters around the world in a standardized form, and thus the target characters include not only currently used scripts but also historic scripts and undeciphered writing systems. With regard to the Uyghur writing system and its characters, the language code of ISO 639 has been registered(Table 1), however, the character name code (ISO 15924) and the character code have not been registered yet (Table 2)[5],[6]. In the roadmap for future standardization published by the expert group involved in the development of ISO/IEC 10646 [7], sign allocation areas have been reserved for ancient writing systems, such as the Uyghur writing system and the Sogdian writing system, which indicates that the importance of such standardization is recognized among experts in character encoding. According to the Roadmap to the SMP (Plane 1) at the time of writing this article, 0D00-0D5F for 96 characters are planned to be allocated to the Uyghur script, 0E00-0E5F for 96 characters to the Sogdian script, and 1F60-1FBF for 96 characters to the Turkestani script. The Turkestani script is one derived from the Brahmi script used for writing Tocharian and Uyghur in around the 8th century.

In the background as said above, we have been engaged in the development of the Uyghur character code. And in order to materialize the registration for ISO/IEC 10646, we informally began to contact ISO/IEC/JTC1/SC2/WG2 and Unicode, and sent the



	0	1	2	3	4	5
100	**Old Turkic**					
101	¿Uighur?					
102	¿Sogdian?					

	6	7	8	9	A	B
11F	¿Turkestani?					

Figure 1: A part of Roadmap to the SMP (Plane 1)[7].

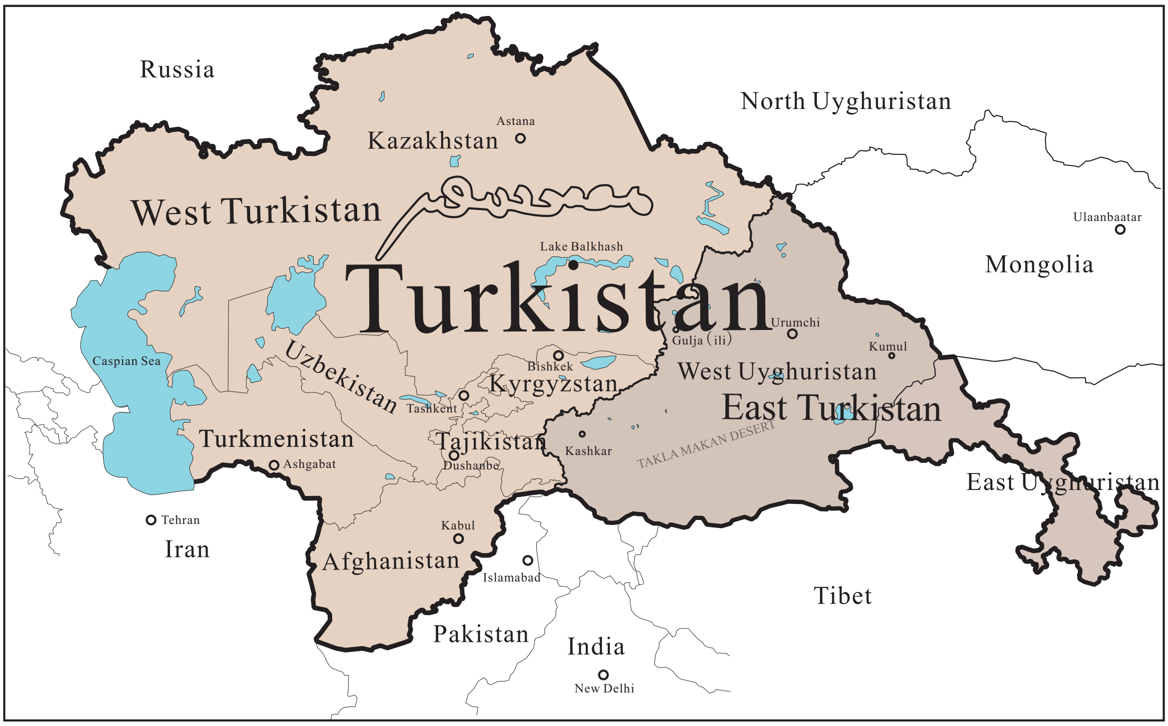


Figure 2: West and East Turkistan[11],([9],p.2,15.line).

first draft of the proposal for Uyghur character encoding. This proposal was improved according to the comments made by experts in standardization, and the proposal described in this article is based on such improvement.

As there are many writing systems and scripts handled in this article, we would like to mention the principles for describing them as follows. For the English marks of the names for languages and scripts, we referred to ISO 639 (Language Code) and ISO 15924 [Code for the representation of names of scripts), and for the names of languages and scripts in English that are not contained in them, we referred to the literature[8].

1.2 About the name of Uyghur

Roadmap to the SMP uses the name of (Uighur)[7]. This name as a base, ISO 639-2 Language code registers (uig)(Table 1). The linguist and the researcher of the Uyghur language are using the name of (Uyghur). Now, the information system of the world is using the name of (Uyghur)[8],[25],[26], [27], [28], [29]. I am using the name of the (Uyghur) in this proposal. I think that registration is necessary for (Uyghur) and (uyg) of Roadmap to the SMP and ISO 639-2 Language Code.

	Non-Enctibil	Enctibil
Vowel	/a/, /æ/, /e/, /i/,	/o/, /ö, / /u/, /ü/.
	Voiced	
Consonant	/b/, /ʤ/, /d/, /r/, /z/, /ʒ/, /ɣ/.	
	/g/, /ŋ/, /l/, /m/, /n/, /v/, /j/.	
	Silent	
	/p/, /t/, /ʧ/, /x/, /s/, /ʃ/, /f/, /q/, /k/, /h/.	

Figure 3: Uyghur vowels and consonants.

1.3 Uyghur Language

The Uyghur people were in Central Asia and North Asia, and their language is called Uyghur (“ Uygur Tili ” in their language). Uyghur is the language belonging to the Turkic (Türk) languages in the Altaic language family. Uyghur has some differences depending on the regions, and there are some differences between the Uyghur languages historically used in different times and modern Uyghur.

The time when the Uyghur script, which is the theme of this article, began to be used was also the time when the Uyghur as a written language was established. The Uyghur writing system at that time was comprised of 8 vowels and 24 consonants. The 8 vowels are the same as those of modern Uyghur; however there are a few differences from the Modern Uyghur, such as a long vowel and a diphthong. Modern Uyghur has 4 rounded vowels and 4 unrounded vowels(Figure 3), however, in the middle Uyghur literature, which was used as the material for the character code design, long vowels such as /uu/ and diphthongs such as /ai/ were separately described[9].

1.4 Uyghur Script

The history of the scripts used by Uyghur people can be classified by the periods when they were used into four groups: Old, Middle, Modern Ages and Modern. The period of Old corresponds to the times up to the end of the 7th century in which Old Turkic and others were used. The period of Middle corresponds to between the 8th and the end of the 19th centuries when the Uyghur script in the broadest sense of the term, including the currently used script and the ancient scripts, refers to the one used in this period[8], [14]. This article handles this script as the theme. The period of Modern Ages corresponds to the end of the 19th century to 1949. During this period, Russia and China made inroads into Turkistan(Figure 2), which led to the use of the Arabic and Cyrillic scripts. The period of Modern corresponds to 1950 up to now. Russia and China made inroads into Turkistan again in 1949, which led to the use of the Latin, Cyrillic, and Arabic scripts. Modern Uyghur script often refers to the Arabic, Cyrillic, and Latin scripts to which new characters were added for the writing of Uyghur, and thus speaking strictly, Modern Uyghur script does not exist.

The Uyghur script handled in this article is a phonogram whose origin is the Aramaic script and direct ancestry is the Sogdian script derived from that. Later the Mongolian, Manchu, and other scripts were derived from the Uyghur script. In this sense, we can say that the Uyghur script is important enough to be regarded as one of the roots of Asian scripts like kanji (Chinese script) and Indian scripts. Figure 4 indicates the historical genealogy in the scripts related to the Uyghur script; specifically, it shows (A) the Aramaic script[12]p.75, (B) Sogdian script[13]p.100, (C) horizontal Uyghur script[9]p.8, (D) vertical Uyghur scripts[2],[3]p.88, (E) Mongolian scripts[8]p.547, (F) Manchu script[8]p.553, (G) Todo (Oirat, Kalmyk) script[8]p.553, and (H) Buryat script[8]p.553 the genealogical relation among them and the rough times when they were used.

The Uyghur script was written both in vertical and horizontal directions. The Uyghur script written in a horizontal direction is seen in West Turkistan and East Turkistan of

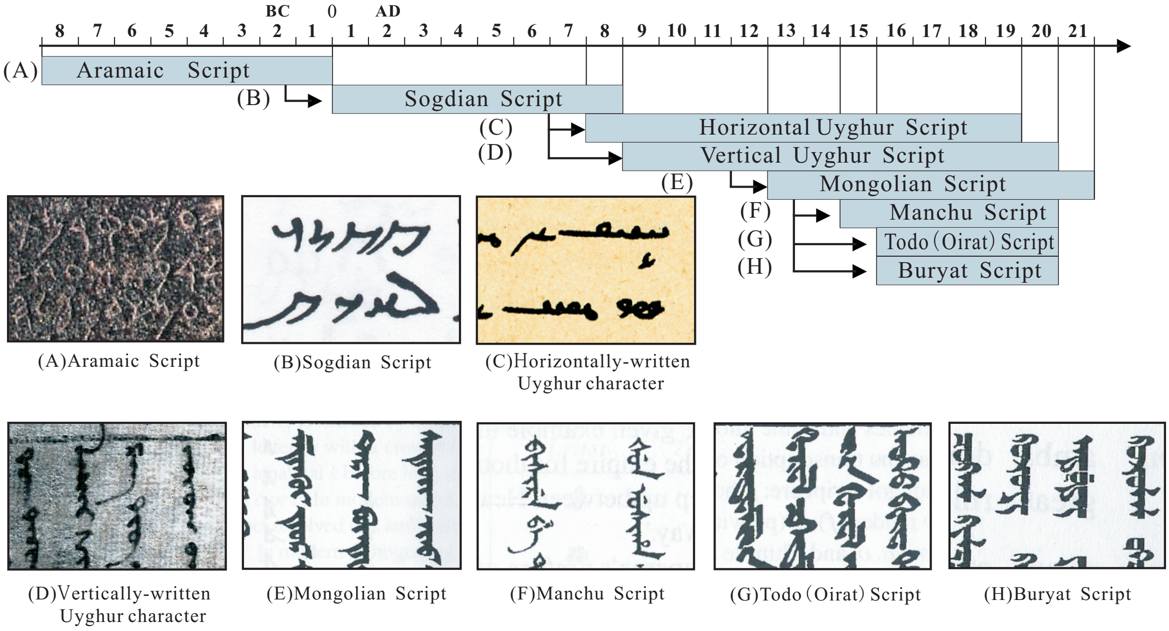
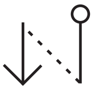
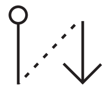
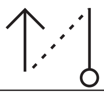
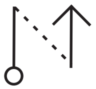
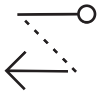
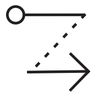


Figure 4: Genealogy in the Uyghur script(A)[12]p.75, (B)[13]p.100, (C)[9]p.8, (D)s[2],[3]p.88, (E)[8]p.547, (F)[8]p.553, (G)[8]p.549, (H)[8]p.554.

Central Asia figure 2. On the other hand, the Uyghur script written in a vertical direction Figure 4 [2],[3] is presumed to have been created around the 8th to 9th century in the Turfan district [2],[3], which came to be used in a wide area from Turfan in the west to Mongolia and Gansu in the east. The horizontal Uyghur and Arabic scripts are written horizontally from right to left, and a new line starts from top to bottom (Table 3(E)). When these scripts are rotated 90 degrees counterclockwise, they are vertically written from top to bottom and a new line starts from left to right just like the Mongolian and Manchu scripts (Table 3(B)). The Uyghur script was changed from horizontal line orientations to vertical line orientations; however, a new line starts the same as that in the scripts in the kanji cultural area (Table 3(A)) such as kanji, kana (the Japanese syllabary), and Hangeul (Table 3(A)). In addition to the Uyghur people, the Mongolian, Kalmyk, Manchu, and Buryat people used the vertical Uyghur script for a literary language [8]. There is a direction of writing from bottom to top Uyghur characters(Table 3(E, F)).

Based on the Uyghur script, the Mongolian script was created (Figure 4 [2],[3](E)) in around the 13th century by the users of Mongolic. The oldest epigraph of the Mongolian script is Genghis Khan ’ s stone inscription around 1225. Between the end of the 16th century and the beginning of the 17th century, the new calligraphy of the Mongolian script was specified, which led to the birth of modern Mongolian script [13]. In the Qing dynasty, which was the dynasty of the Manchu people, Nurhachi (Founder null wasp) ordered his vassals to create the Manchu script newly based on the Mongolian script in 1599(Figure 4(F))[13]. The Todo (Oirat, Kalmyk) script was created in 1648 and was based on the Mongolian script (Figure 4(G))[13]. Later, the Buryat script (Figure 4(H)) was created by users of the Buryat language.

Table 3: Writing Direction of the documents written in Various scripts.

A Vertically written left line	B Vertically written right line	C Vertically written left line	D Vertically written right line	E Horizontal writing for the left	F Horizontal writing for the right
					
Uyghur script Kana Hangeul Kanji	Uyghur script Mongol script Manchu script Todo script Buryat script	Uyghur script	Uyghur script	Uyghur script Arabic script	Latin script Kana Hangeul Kanji

1.5 Materials Used in This propose as Sources

In this article, we referred to Kutadgu Bilig“ Wisdom of Royal Glory ”as the text of the horizontal Uyghur script [9]. It was published by Turk dil kurumu (The Turkish Language Association) in istanbul in 1942, but the photography and printing of the original were done by C. Angerer and Goschl, the famous printing office in Vienna. Kutadgu Bilig is a work created by Yusuf Khass Hajib who was an 11th century Uyghur literary man [15], [16], [17], which is one of the precious cultural assets of Turkestan. Some scholars say that the role of Kutadgu Bilig in the history of the Uyghur language may correspond to that of La Divina Commedia (The Divine Comedy) by Dante in the Italian language [15], and thus it is a literature with a significant meaning for not only scholars in the Uyghur language butand also the Uyghur people. The time when Kutadgu Bilig was written is not described in this transcript itself, however, Turk dil kurumu which published the photographic reprint of the copy in the 20th century says that it was written in 1069 or 1070 [9]. For Kutadgu Bilig, not only the Vienna transcript I referred to for writing this article (Viyana Nushasi, owned by Vienna King Library [9] but the transcript written in the Arabic script is known. In addition, the Fergana transcript (Gergana Nushasi) and the Egyptian transcript (Misir Nushasi) are also known now [18], [19].

On the other hand, the text of the vertical Uyghur script used for reference purposes in this article is “ Abhidharmakosabhasyatika Tattvartha. ” This was obtained in East Turkestan by an Englishman Marc Aurel Stein in 1907. It is archived in the British Library now, and Or8212-75A and Or8212-75B are allocated to it as the reference numbers in the library. S, .Tekin who is the researcher of ancient Turkish said that it was written in around 1300 to 1400 [20]. In this article, the photocopy material in the writing of Masahiro Shogaito [3] is used.

2 Design of the Uyghur Character Code

2.1 Principles of Character Code Design

We researched the principles of character code design required for designing the Uyghur character code. The base is ASCII (American Standard Code for Information Interchange), which is the starting point for the character code. Gorn, Bemer, and Green, who were the designers of ASCII, mentioned the proposal in 1963 that there were principles in the design of character codes as follows [21];

(1) Preparing the appropriate number of graphics necessary for symbolizing characters.

(2) Preparing the appropriate number of codes necessary for control.
(3) Eliminating ambiguity.
(4) Restrictions related to media and devices.
(5) Function for controlling (correcting) errors.
(6) Special interpretation of code for which all the bits are 0 (or 1) (NULL and DEL).
(7) Easy identification of character classes.
(8) Convenience in data processing (Easy case conversion).
(9) Order of array (logical, historical).
(10) Keyboard array (logical, historical).
(11) Other size factors.
(12) Easy internationalization.
(13) Ability to write a programming language.
(14) Compatibility with existing codes.

Among the above, the matters of control code in (2), NULL and DELL in (6), a programing language in (13), and the consistency with a keyboard array in (10) have already been solved considering that ISO/IEC 10646 is backward compatible with ASCII. The matter of case conversion in (7) is also excluded because it is not necessary for the Uyghur script. Since the character code that we try to design this time is not for a single table but for a part to be added to ISO/IEC 10646, there is no restriction on the seven or eight-bit range that was the pattern of character code design in the early stage. Thus, the matters of (4) and (11) may be excluded. Though ISO/IEC 10646 does not have the function of error correction, an encoding that can detect errors, such as UTF-8, is defined. It is available for the zones to which no character code is allocated, thus we can say that the matter in (5) has been solved. As (12) is set considering the case where ASCII is used for languages using the Latin script other than English, this matter may be excluded. And since the Uyghur script has no existing code, the matter in (14) may be excluded, too.

Therefore, the four principles are left as follows:
(1) Preparing the appropriate number of graphics necessary for symbolizing characters.
(3) Eliminating ambiguity.
(7) Easy identification of character classes.
(9) Order of sequence (logical, historical).

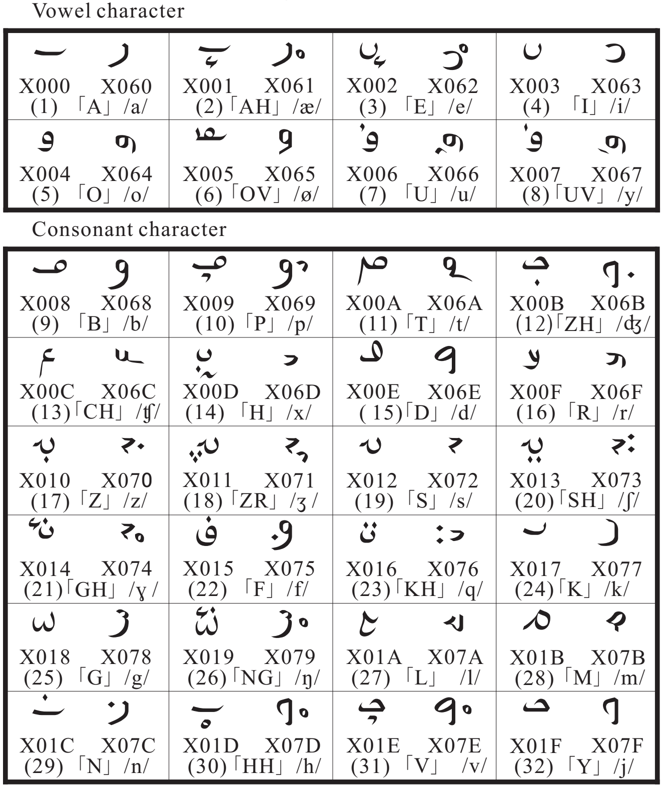
2.2 Application of the principles of the design to the Uyghur script

Now we discuss by applying specifically the said four principles to the Uyghur script.

First, we examine how the scope of the Uyghur characters included should be determined according to the principles (1) and (3). Kutadgu Bilig and Abhidharmakosabhasyatika Tattvartha for the examination have various figures that appear to indicate the signs of authors, and the number of the codes will, if we encode all of them, reach several hundreds. Therefore, in designing the character code, we determined the scope of the characters from the viewpoint of the character set that should be required for representing Uyghur as a phonographic writing system.

Now the matter to be considered here is whether it should be necessary to distinguish a code between the characters for vertical writing and those for horizontal writing. The similar examples of character sets written both vertically and horizontally that are already encoded in ISO/IEC 10646 are kanji, Hangul, and kana (the Japanese syllabary); however, none of these scripts are given a character code differently between vertical writing and horizontal writing. For the Uyghur script, however, it is necessary to rotate the characters according to the text direction because the Uyghur characters are connected to each other when written. In addition, as shown in table 9, the changes in the letterforms that cannot be handled by simple rotation are seen in many Uyghur characters. This aspect is specific to the Uyghur script, which is different from kanji, Hangul, or kana after type came to be used, and thus should be considered in encoding the Uyghur script. The

Table 4: Uyghur Character Set.



Vowel character

X000 X060 (1) 「A」 /a/	X001 X061 (2)「AH」 /æ/	X002 X062 (3) 「E」 /e/	X003 X063 (4) 「I」 /i/
X004 X064 (5) 「O」 /o/	X005 X065 (6)「OV」 /ø/	X006 X066 (7) 「U」 /u/	X007 X067 (8)「UV」 /y/

Consonant character

X008 X068 (9) 「B」 /b/	X009 X069 (10)「P」 /p/	X00A X06A (11)「T」 /t/	X00B X06B (12)「ZH」 /ʤ/
X00C X06C (13)「CH」 /ʧ/	X00D X06D (14) 「H」 /x/	X00E X06E (15)「D」 /d/	X00F X06F (16) 「R」 /r/
X010 X070 (17) 「Z」 /z/	X011 X071 (18)「ZR」 /ʒ/	X012 X072 (19) 「S」 /s/	X013 X073 (20)「SH」 /ʃ/
X014 X074 (21)「GH」 /ɣ/	X015 X075 (22) 「F」 /f/	X016 X076 (23)「KH」 /q/	X017 X077 (24)「K」 /k/
X018 X078 (25) 「G」 /g/	X019 X079 (26)「NG」 /ŋ/	X01A X07A (27) 「L」 /l/	X01B X07B (28) 「M」 /m/
X01C X07C (29) 「N」 /n/	X01D X07D (30)「HH」 /h/	X01E X07E (31) 「V」 /v/	X01F X07F (32) 「Y」 /j/

quotation marks and punctuation marks used in kana have to change their positions and orientations according to the text direction, but different codes are not given to them except for the characters for compatibility with the character code standard before ISO/IEC 10646, and they are processed in the form of selecting an appropriate glyph at the time of output. In the Uyghur script, many of the characters have differences in their forms between the vertical writing and horizontal writing, which cannot be handled by simple rotation; however, on the other hand, separating the vertically written Uyghur character code from that of the horizontal one has both merits and demerits considering that mixed use of vertical writing and horizontal writing is highly unlikely, and it would be convenient to use the code point common to both of them in character retrieval. Finally, we referred the decision on this matter to ISO/IEC for discussions on international standardization, and we created the character codes for both the vertical and horizontal scripts in this study.

Another important matter is the discrimination between glyphs and codes. This matter was not recognized at the time of ASCII, but in order to handle scripts whose letterforms change depending on the positions in words, such as Arabic or Uyghur, the following schemes are used: the scheme where a different code is given to each different figure, and the scheme where the same code is given to the figures whose shapes are different but have the same phoneme (code-glyph separation scheme). As the code-glyph separation scheme is adopted in ISO/IEC 10646, we also adopted it into the design this time. Code positions are assigned to some presentation forms of the Arabic script for the purpose of backward compatibility with the character code before ISO/IEC 10646. However, as we expected that the principle of character-glyph separation would be strictly applied to the proposal of new standardization, we designed the character code as such.

The matter of identification of character classes in (7) needs to be considered in the Uyghur script. In ASCII, the task was to identify three character classes, i.e. letters, numerals, and marks (as lower-case letters were added later, it became necessary to identify four character classes). For the Uyghur script, different from the Latin script, it

Table 5: Uyghur character classification.

Category	Category Name	Character Symbol	
		Horizontal	Vertical
V	Vowels	HV（00～07）	VV（60～67）
C	Consonants	HC（08～1F）	VC（68～7F）
D	Diacritical Marks	HD（20～2C）	VD（80～87）
P	Punctuation Marks	HP（30～36）	VP（90～97）
N	Numerals	HN（40～49）	VN（A0～A9）
J	Signature Symbols	HJ（50）	VJ（B0～BD）

is necessary to distinguish vowels, consonants, phonetic symbols, and others. Thus, as shown in table 4, it is necessary to identify 12 classes if vertical writing is distinguished from horizontal writing and 6 classes if not. And these classes have to be allocated in the definite blocks mapped in the code table.

The last principle - order of array in (9) means that if the order of character sequence has been historically defined, the code table should correspond to it as much as possible. However, there is no information about the order of array in the Uyghur script at that time, so we adopted the same order as that of the Modern Uyghur script.

2.3 Character Set

The abbreviations used for representing the vertical and horizontal Uyghur script are as follows.

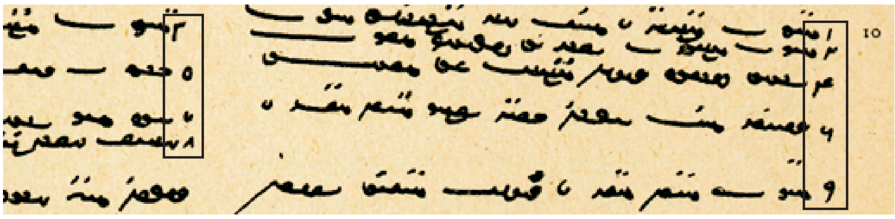
In this article, vowels shall be represented as V, consonants as C, diacritics marks as D, punctuation marks as P, numerals as N, and signature symbols as S, and abbreviations V or H, which indicates whether the Uyghur character set is for vertical writing or horizontal writing (Table 4), shall be added to the heads of those six abbreviations for identifying the characters. The code tables for the overall Uyghur script represented as above are shown in (Table 9).

The vertical and horizontal Uyghur writing system is comprised of phonograms representing vowels and consonants. Thus the character sets are for 8 horizontal vowel letters and 8 vertical vowel letters corresponding to 8 vowels, and for 24 horizontal consonant letters and 24 vertical consonant letters corresponding to 24 consonants (Table 4). New scripts, when they are added in ISO/IEC 10646, are usually allocated to the code points in 8-bit boundaries first, and thus the starting point shall be X000, and the code point of each character proposed shall be expressed as the offset from the starting point. The number inside the parentheses in the third line indicates the serial number of the Uyghur alphabet written vertically and horizontally, and within // is the pronunciation using the symbol from IPA. Since the encoded characters in ISO/IEC 10646 are given the names comprised of ASCII only (numbers and uppercase Roman alphabet), we showed the pronunciation in ASCII characters in 「」. The letters in 「」 are the pronunciations of horizontal and vertical Uyghur script that we propose as the names of the Uyghur characters written horizontally and vertically. The sets of the Uyghur characters comprise the vowels, consonants, diacritics marks, punctuation marks, numerals, and signature symbols.

2.4 Number of characters

Uyghur numbers is classified into two kinds of Round digits and Lattice digits.

Round digits is used horizontally written from right to left. Round digits is written



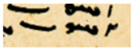
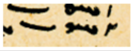
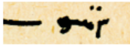
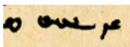
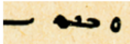
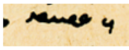
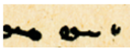
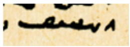
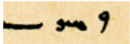
No	Numbers	Sample image	Name of numbers	Reference
0	٠		UYGHUR ROUND DIGIT ZERO	
1	١		UYGHUR ROUND DIGIT ONE	Line 10
2	٢		UYGHUR ROUND DIGIT TWO	Line 11
3	٣		UYGHUR ROUND DIGIT THREE	Line 10
4	۴		UYGHUR ROUND DIGIT FOUR	Line 12
5	٥		UYGHUR ROUND DIGIT FIVE	Line 12
6	٦		UYGHUR ROUND DIGIT SIX	Line 13
7	٧		UYGHUR ROUND DIGIT SEVEN	Line 13
8	٨		UYGHUR ROUND DIGIT EIGHT	Line 13
9	٩		UYGHUR ROUND DIGIT NINE	Line 14

Figure 5: Round numbers of Uyghur characters[9]124.p.

from the left to the right with the document of Horizontally written from right to left. It is similar to the written Latin digits for the writing of Round Uyghur digits.

Name of numbers of Round digits is as follows(Figure 5).

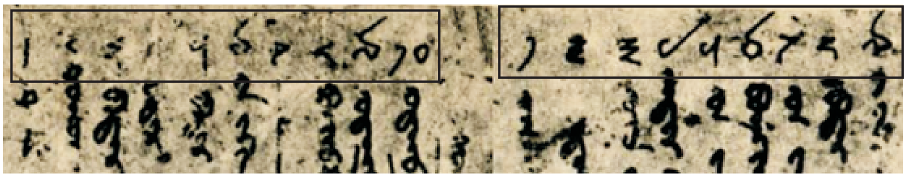
0. UYGHUR ROUND DIGIT ZERO
1. UYGHUR ROUND DIGIT ONE
2. UYGHUR ROUND DIGIT TWO
3. UYGHUR ROUND DIGIT THREE
4. UYGHUR ROUND DIGIT FOUR
5. UYGHUR ROUND DIGIT FIVE
6. UYGHUR ROUND DIGIT SIX
7. UYGHUR ROUND DIGIT SEVEN
8. UYGHUR ROUND DIGIT EIGHT
9. UYGHUR ROUND DIGIT NINE

Lattice digits is used Vertically written from top to bottom and bottom to top. Round digits is written from the left to the right and from top to bottom with the document of Vertically written from from top to bottom and bottom to top.

It is similar to the written Latin digits and it is the same as Japanese vertically written digits for the writing.

Name of numbers of lattice digits is as follows(Figure 6).

0. UYGHUR LATTICE DIGIT ZERO
1. UYGHUR LATTICE DIGIT ONE
2. UYGHUR LATTICE DIGIT TWO
3. UYGHUR LATTICE DIGIT THREE
4. UYGHUR LATTICE DIGIT FOUR
5. UYGHUR LATTICE DIGIT FIVE
6. UYGHUR LATTICE DIGIT SIX
7. UYGHUR LATTICE DIGIT SEVEN



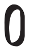
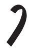
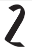
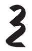
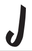
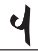
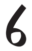
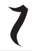
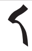
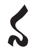
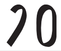
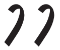
No	Numbers	Sample image	Name of numbers	Reference
0			UYGHUR LATTICE DIGIT ZERO	Line 16
1			UYGHUR LATTICE DIGIT ONE	Line 20
2			UYGHUR LATTICE DIGIT TWO	Line 21
3			UYGHUR LATTICE DIGIT THREE	Line 22
4			UYGHUR LATTICE DIGIT FOUR	Line 23
5			UYGHUR LATTICE DIGIT FIVE	Line 24
6			UYGHUR LATTICE DIGIT SIX	Line 25
7			UYGHUR LATTICE DIGIT SEVEN	Line 26
8			UYGHUR LATTICE DIGIT EIGHT	Line 14
9			UYGHUR LATTICE DIGIT NINE	Line 15
10				Line 16
10				

Figure 6: Lattice numbers of Uyghur characters[10].

8. UYGHUR LATTICE DIGIT EIGHT
9. UYGHUR LATTICE DIGIT NINE

2.5 Names of the characters

In the history, The culture, Religion of the Uyghur people who lives in west Asia, north Asia, Central Asia, and east Asia it is different. Uyghur that lives in west Asia is called the west Uyghur people. They, characters used, are called the west Uyghur character. The west Uyghur character is writing from right to left. However, digits is written from left to right direction just like the Latin digits.

Uyghur that lives in central Asia is called the east Uyghur people. They, characters used, are called the east Uyghur character. The east Uyghur character is writing from top to bottom and bottom to top. Digits is written from top to bottom direction just like the Japanese digits.

The name of the west Uyghur character is as follows(Table 9).

West Uyghur vowel letter

X000 WEST UYGHUR VOWEL LETTER A
X001 WEST UYGHUR VOWEL LETTER AH
X002 WEST UYGHUR VOWEL LETTER E
X003 WEST UYGHUR VOWEL LETTER I
X004 WEST UYGHUR VOWEL LETTER O
X005 WEST UYGHUR VOWEL LETTER OV
X006 WEST UYGHUR VOWEL LETTER U

X007 WEST UYGHUR VOWEL LETTER UV

West Uyghur consonant letter

X008 WEST UYGHUR CONSONANT LETTER B
X009 WEST UYGHUR CONSONANT LETTER P
X00A WEST UYGHUR CONSONANT LETTER T
X00B WEST UYGHUR CONSONANT LETTER ZH
X00C WEST UYGHUR CONSONANT LETTER CH
X00D WEST UYGHUR CONSONANT LETTER H
X00E WEST UYGHUR CONSONANT LETTER D
X00F WEST UYGHUR CONSONANT LETTER R
X010 WEST UYGHUR CONSONANT LETTER Z
X011 WEST UYGHUR CONSONANT LETTER ZHE
X012 WEST UYGHUR CONSONANT LETTER S
X013 WEST UYGHUR CONSONANT LETTER SH
X014 WEST UYGHUR CONSONANT LETTER GK
X015 WEST UYGHUR CONSONANT LETTER F
X016 WEST UYGHUR CONSONANT LETTER KH
X017 WEST UYGHUR CONSONANT LETTER K
X018 WEST UYGHUR CONSONANT LETTER G
X019 WEST UYGHUR CONSONANT LETTER NG
X01A WEST UYGHUR CONSONANT LETTER L
X01B WEST UYGHUR CONSONANT LETTER M
X01C WEST UYGHUR CONSONANT LETTER N
X01D WEST UYGHUR CONSONANT LETTER HH
X01E WEST UYGHUR CONSONANT LETTER V
X01F WEST UYGHUR CONSONANT LETTER Y

West Uyghur diacritics marks

X020 WEST UYGHUR DIACRITICS MARKS HAMZA ABOVE
X021 WEST UYGHUR DIACRITICS MARKS HAMZA BELOW
X022 WEST UYGHUR DIACRITICS MARKS HAMZA DOT
X023 WEST UYGHUR DIACRITICS MARKS U PASH
X024 WEST UYGHUR DIACRITICS MARKS UV PASH
X025 WEST UYGHUR DIACRITICS MARKS PASH BELOW
X026 WEST UYGHUR DIACRITICS MARKS DOT ABOVE
X027 WEST UYGHUR DIACRITICS MARKS DOT BELOW
X028 WEST UYGHUR DIACRITICS MARKS TWO DOTS ABOVE
X029 WEST UYGHUR DIACRITICS MARKS TWO DOTS BELOW
X02A WEST UYGHUR DIACRITICS MARKS THREE DOTS
X02B WEST UYGHUR DIACRITICS MARKS KASH AGMA DOT
X02C WEST UYGHUR DIACRITICS MARKS EARRINGS

West Uyghur punctuation symbol

X030 WEST UYGHUR PUNCTUATION SYMBOL MIDDLE DOT
X031 WEST UYGHUR PUNCTUATION SYMBOL STRAIGHT LINE
X032 WEST UYGHUR PUNCTUATION SYMBOL OBLIQUE LINE
X033 WEST UYGHUR PUNCTUATION SYMBOL VERTICAL OBLIQUE LINE
X034 WEST UYGHUR PUNCTUATION SYMBOL LEFT COMMA
X035 WEST UYGHUR PUNCTUATION SYMBOL RIGHT COMMA
X036 WEST UYGHUR PUNCTUATION SYMBOL HYPERBOLA

West Uyghur digits

X040 WEST UYGHUR DIGIT ZERO
X041 WEST UYGHUR DIGIT ONE
X042 WEST UYGHUR DIGIT TWO
X043 WEST UYGHUR DIGIT THREE
X044 WEST UYGHUR DIGIT FOUR

X045 WEST UYGHUR DIGIT FIVE
X046 WEST UYGHUR DIGIT SIX
X047 WEST UYGHUR DIGIT SEVEN
X048 WEST UYGHUR DIGIT EIGHT
X049 WEST UYGHUR DIGIT NINE

West Uyghur signature symbol
X050 WEST UYGHUR SIGNATURE SYMBOL ARA

The name of the east Uyghur character is as follows(Table 9).

East Uyghur vowel letter
X060 EAST UYGHUR VOWEL LETTER A
X061 EAST UYGHUR VOWEL LETTER AH
X062 EAST UYGHUR VOWEL LETTER E
X063 EAST UYGHUR VOWEL LETTER I
X064 EAST UYGHUR VOWEL LETTER O
X065 EAST UYGHUR VOWEL LETTER OV
X066 EAST UYGHUR VOWEL LETTER U
X067 EAST UYGHUR VOWEL LETTER UV

East Uyghur consonant letter
X068 EAST UYGHUR CONSONANT LETTER B
X069 EAST UYGHUR CONSONANT LETTER P
X06A EAST UYGHUR CONSONANT LETTER T
X06B EAST UYGHUR CONSONANT LETTER ZH
X06C EAST UYGHUR CONSONANT LETTER CH
X06D EAST UYGHUR CONSONANT LETTER H
X06E EAST UYGHUR CONSONANT LETTER D
X06F EAST UYGHUR CONSONANT LETTER R
X070 EAST UYGHUR CONSONANT LETTER Z
X071 EAST UYGHUR CONSONANT LETTER ZHE
X072 EAST UYGHUR CONSONANT LETTER S
X073 EAST UYGHUR CONSONANT LETTER SH
X074 EAST UYGHUR CONSONANT LETTER GK
X075 EAST UYGHUR CONSONANT LETTER F
X076 EAST UYGHUR CONSONANT LETTER KH
X077 EAST UYGHUR CONSONANT LETTER K
X078 EAST UYGHUR CONSONANT LETTER G
X079 EAST UYGHUR CONSONANT LETTER NG
X07A EAST UYGHUR CONSONANT LETTER L
X07B EAST UYGHUR CONSONANT LETTER M
X07C EAST UYGHUR CONSONANT LETTER N
X07D EAST UYGHUR CONSONANT LETTER HH
X07E EAST UYGHUR CONSONANT LETTER V
X07F EAST UYGHUR CONSONANT LETTER Y

East Uyghur diacritics marks
X080 EAST UYGHUR DIACRITICS MARKS EARRINGS
X081 EAST UYGHUR DIACRITICS MARKS U PASH
X082 EAST UYGHUR DIACRITICS MARKS UV PASH
X083 EAST UYGHUR DIACRITICS MARKS PASH
X084 EAST UYGHUR DIACRITICS MARKS DOT RIGHT
X085 EAST UYGHUR DIACRITICS MARKS DOT LEFT
X086 EAST UYGHUR DIACRITICS MARKS TWO DOTS RIGHT
X087 EAST UYGHUR DIACRITICS MARKS TWO DOTS LEFT

East Uyghur pungtuation symbol

X090 EAST UYGHUR PUNCTUATION SYMBOL MIDDLE DOT
X091 EAST UYGHUR PUNCTUATION SYMBOL VERTICAL TWO DOTS
X092 EAST UYGHUR PUNCTUATION SYMBOL FULL STOP
X093 EAST UYGHUR PUNCTUATION SYMBOL VERTICAL THREE DOTS
X094 EAST UYGHUR PUNCTUATION SYMBOL HORIZONTAL THREE DOTS
X095 EAST UYGHUR PUNCTUATION SYMBOL END OF PARAGRAPH
X096 EAST UYGHUR PUNCTUATION SYMBOL FIVE DOTS
X097 EAST UYGHUR PUNCTUATION SYMBOL SIX DOTS

East Uyghur digits
X0A0 EAST UYGHUR DIGIT ZERO
X0A1 EAST UYGHUR DIGIT ONE
X0A2 EAST UYGHUR DIGIT TWO
X0A3 EAST UYGHUR DIGIT THREE
X0A4 EAST UYGHUR DIGIT FOUR
X0A5 EAST UYGHUR DIGIT FIVE
X0A6 EAST UYGHUR DIGIT SIX
X0A7 EAST UYGHUR DIGIT SEVEN
X0A8 EAST UYGHUR DIGIT EIGHT
X0A9 EAST UYGHUR DIGIT NINE

East Uyghur signature symbol
X0B0 EAST UYGHUR SIGNATURE SYMBOL SPIRAL
X0B1 EAST UYGHUR SIGNATURE SYMBOL MANY SPIRAL
X0B2 EAST UYGHUR SIGNATURE SYMBOL SAND TABLE
X0B3 EAST UYGHUR SIGNATURE SYMBOL LEAF
X0B4 EAST UYGHUR SIGNATURE SYMBOL RING
X0B5 EAST UYGHUR SIGNATURE SYMBOL FOUR LINE
X0B6 EAST UYGHUR SIGNATURE SYMBOL PLUS SIGN
X0B7 EAST UYGHUR SIGNATURE SYMBOL UNLIMITED PLUS SIGN
X0B8 EAST UYGHUR SIGNATURE SYMBOL STAR
X0B9 EAST UYGHUR SIGNATURE SYMBOL SIX STAR
X0BA EAST UYGHUR SIGNATURE SYMBOL EIGHT STAR
X0AB EAST UYGHUR SIGNATURE SYMBOL EIGHT STAR DOT

3 Design of the Glyph Table for the Uyghur Script

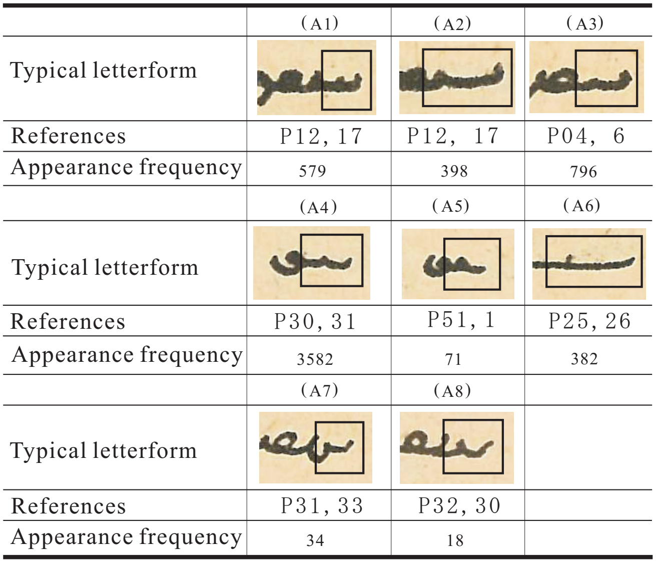
3.1 Procedure for Glyph Design

A glyph is “ an abstract form of character that represents a letterform ” [22], and the glyph design means the task of selecting typical letterforms through careful examinations of characters appearing in the authority materials, and designing abstract forms of those characters based on the examinations. The abstract letterforms only refer to the skeleton outlines of line drawing that do not take into account the width of the lines and others; however, as they will be used practically as a font for output indication, we created them as an outline font designed in line with the change in length and width, angles of the bend sections, and the curvatures, etc., in the line drawing of the letterforms appearing in the authority materials as much as possible.

The units for designing glyphs were presentation forms. Based on the “ code-glyph separation principle ” mentioned in 2.2, the same codes were assigned to the characters irrespective of the difference in the presentation forms; however, the glyphs were created for each of the four presentation forms. In addition, as it was necessary to create glyphs for each of horizontal writing and vertical writing differently, we designed 8 glyphs in total for each character.

Omarjan Osman, one of the authors of this article, is a specialist in the Old Uyghur script, and his mother language is Uyghur. He can read the Old, Middle, and Modern

Table 6: Shake of letterform.



	(A1)	(A2)	(A3)
Typical letterform			
References	P12, 17	P12, 17	P04, 6
Appearance frequency	579	398	796
	(A4)	(A5)	(A6)
Typical letterform			
References	P30, 31	P51, 1	P25, 26
Appearance frequency	3582	71	382
	(A7)	(A8)	
Typical letterform			
References	P31, 33	P32, 30	
Appearance frequency	34	18	

Uyghur script, horizontally and vertically. The detailed procedure of glyph design performed by him was as follows.

With regard to the horizontal Uyghur script, he read from the beginning the text of Kutadgu Bilig, which was the authority material for character code design, and the letterforms that were judged to be new ones were cut out into a JPEG file as “typical letterforms,” which were recorded together with the pages and lines where they appeared. The letterforms cut out were in the form of a character string with a certain length, including the characters before and after the letterform, in order to save the information, because the letterforms are affected by the characters before and after them. The occurrence count in which the judged letterforms were identified as the same as any of the typical ones that had already appeared was totaled. Thus, the text was scanned to the last page, and the outline font was designed based on the typical letterforms that appeared the most frequently.

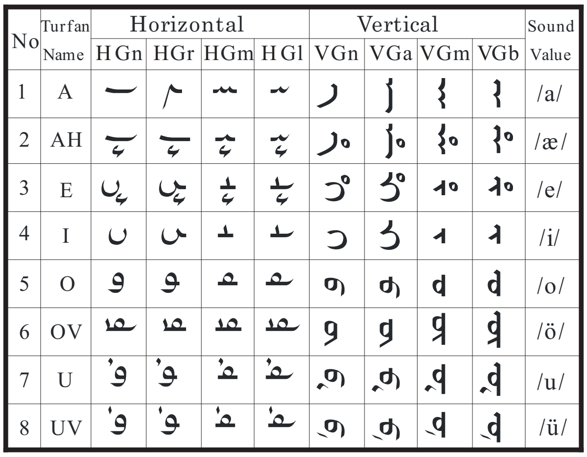
As an example, the shake of the form of the Uyghur character representing the pronunciation /a/ is shown in Table 6. This character is composed of two upward projections and horizontal lines connecting the projections, and subtle differences were found in the length of the gap between the projections and the height and/or the angle of the projections. In this example, 8 typical letterforms were extracted, each of which appeared 579 times, 398 times, 796 times, 3582 times, 60 times, 340 times, 22 times, and 16 times, respectively.

As this task demanded high concentration because subtle differences in letterforms should be identified properly, he had to focus on the particular presentation forms of the particular characters. He ended up scanning the text 128 times (32 characters x 4 presentation forms) repeatedly from its beginning to the end, and it took him around 1500 hours. Through this examination, 398 typical letterforms were extracted in total, that is, 3.1 typical letterforms per presentation form (= 398/32 characters x 4 presentation forms) were extracted on average. In addition to vowels (8 characters x 4 presentation forms = 32 glyphs) and consonants (24 characters x 4 presentation forms = 96 glyphs), the glyphs of diacritics marks (13 glyphs), punctuation marks (7 glyphs), numerals (10 glyphs), and a signature symbol (1 glyph) were also designed. Thus, the total number of glyphs designed was 159.

With regard to the vertical Uyghur script, the same examination was carried out using Abhidharmakosabhasyatika Tattvartha as the authority material. In this examination, the text of the authority material was scanned from the beginning to the last page (page 501), and 215 typical letterforms were extracted in total. This task took the author

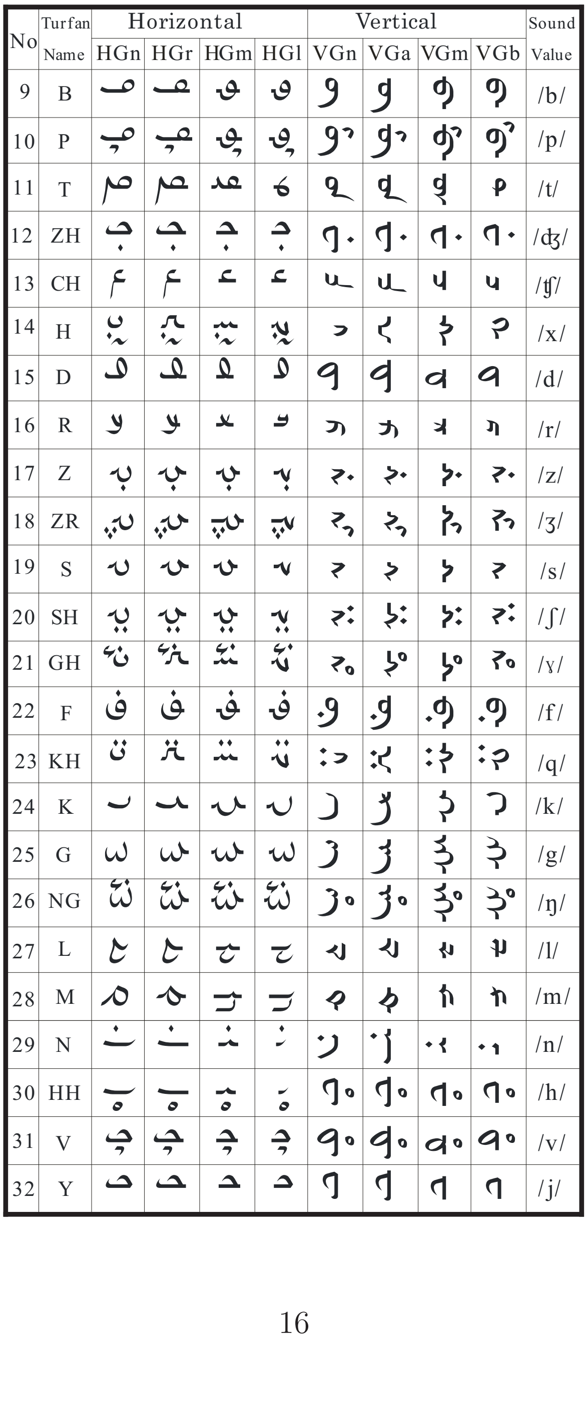
Table 7: Uyghur glyphs.

Vowel glyph



No	Turfan Name	Sound Value
1	A	/a/
2	AH	/æ/
3	E	/e/
4	I	/i/
5	O	/o/
6	OV	/ö/
7	U	/u/
8	UV	/ü/

Consonant glyph



No	Turfan Name	Sound Value
9	B	/b/
10	P	/p/
11	T	/t/
12	ZH	/ʤ/
13	CH	/ʧ/
14	H	/x/
15	D	/d/
16	R	/r/
17	Z	/z/
18	ZR	/ʒ/
19	S	/s/
20	SH	/ʃ/
21	GH	/ɣ/
22	F	/f/
23	KH	/q/
24	K	/k/
25	G	/g/
26	NG	/ŋ/
27	L	/l/
28	M	/m/
29	N	/n/
30	HH	/h/
31	V	/v/
32	Y	/j/

around 730 hours. The total number of glyphs for the vertical script was 166; 8 characters x 4 presentation forms = 32 glyphs for vowels, 24 characters x 4 presentation forms = 96 glyphs for consonant, 8 glyphs for diacritics marks, 8 glyphs for punctuation marks, 10 glyphs for numerals, and 12 glyphs for signature symbols.

The font creation tool and the environment used for the said task were as follows;

- Design: $Adobe^{\circledR}Photoshop^{\circledR}CS5ExtendedandAutodesk^{\circledR}3dsMax^{\circledR}9$
- Conversion into an outline font: Font Creater 5.6
- Font file type: TrueType font (ttf)

3.2 Glyph Table

The designed glyphs were allocated to the table with the axes of the corresponding character code and presentation form, respectively, and each glyph was identified by four alphanumeric characters consisting of the symbol indicating whether it is for vertical writing or horizontal writing (H|V), the symbol indicating that it is a glyph (G), the symbol indicating the presentation forms (l|m|r|n for horizontal writing, and b|m|a|n for vertical writing), and the figure indicating the character. The l|m|r|n are the abbreviations for left-joining, medial-joining, right-joining, and nominal, respectively, and the b|m|a|n are for below-joining, medial-joining, above-joining, and nominal, respectively. For example, HGn2 means the glyph of the second character used in the nominal form for horizontal writing, and VGa28 means the glyph of the 28th character used in the above-joining form for vertical writing. This glyph table created for horizontal writing was sent, as the first proposal for the Uyghur character code, to the experts of ISO/IEC/JTC1/SC2/WG2 and the Unicode Technical Committee in 2008, together with the JPEG file of the typical letterforms from which the glyphs were derived and the data on the pages and lines where the letterforms appear in the authority material. And, the glyph table created for vertical writing was sent to the experts of ISO/IEC/JTC1/SC2/WG2 and the Unicode Technical Committee in 2011, together with the JPEG file of the typical letterforms from which the glyphs were derived and the data on the pages and lines where the letterforms appear in the authority material. Table 10 shows the at-a-glance glyph table for both the horizontal and vertical writing.

4 Verification

4.1 Purpose and Method of Verification

We conducted verification about whether the character code table and the glyph set for horizontal and vertical writing that were created in this study would be good for actual usage, and whether they were aligned with the principles adopted in this encoding, according to the following method.

First, we verified whether this glyph set could cover the graphics necessary for describing Kutadgu Bilig and Abhidharmakosabhasyatika Tattvartha. This verification is for the principle of figure set securing (2.1 (1) of this article). We created the glyphs very carefully as mentioned in 3.1; however, it is still necessary to check whether there is any oversight in the coverage of graphics. We verified that matter by reproducing the texts of Kutadgu Bilig and Abhidharmakosabhasyatika Tattvartha using the glyph set we designed. Normally, it is preferable that all the texts are input for this verification, however, the full texts of Kutadgu Bilig and Abhidharmakosabhasyatika Tattvartha are extremely huge in volume (around 79,600 characters in 199 pages and around 100,200 characters in 501 pages respectively), and inputting the full texts without appropriate input support software would take a good amount of time. Therefore, we implemented the input experiment with the texts in the pages from the beginning to page 12 of both Kutadgu Bilig and Abhidharmakosabhasyatika Tattvartha. The number of characters input resulted in 3,787 from Kutadgu Bilig and 1661 from Abhidharmakosabhasyatika Tattvartha. If there is any glyph that is necessary for the texts but impossible to input, it means that the coverage of the glyph set is not sufficient. As a result of this verification experiment, there was no lack of glyphs found.

Though this is the verification of part of the texts, we think it may be sufficient for the verification of the coverage because of the following reasons.

It is generally said that the appearance frequency of a certain character or word in natural languages roughly follows Zipf's law. Zipf's law is an empirical rule that the appearance probability Pn of the character in frequency rank n is inversely proportional to the reciprocal of n. As the number of glyph types of the Uyghur script is around 128 as shown in Table 7, when the equation $Pn = k/n$is set up using the constant k, k can be evaluated by the following equations.

$$\sum_{n=1}^{128} P_n = \sum_{n=1}^{128} k/n = 1$$

As k 1/4.85, the appearance probability of the glyph which appears the least frequently P_{128} is $k/128$ 0.0016, and thus the probability Q_{3787} that this glyph never appears in the text consisting of 3787 characters is around 0.0022, and the probability Q_{1661} that this glyph never appears in the text consisting of 1661 characters is around 0.00687. Since the probability of oversight may still remain in those texts at 0.2 %, and 6.9 % respectively, we would like continue the verification in the future for perfection. Figure 5 shows the distribution of occurrence probabilities of character code as a unit in the texts input. Actual distributions are represented by the curves descending more sharply than the Zipf law. Considering this result, the probability of oversight may be a little higher.

$$Q_{3787} = \left(1 - \frac{1}{4.85 * 128}\right)^{3787} = 0.0022$$

$$Q_{1661} = \left(1 - \frac{1}{4.85 * 128}\right)^{1661} = 0.0687$$

Second, we verified whether more than one glyph would correspond to one phoneme, or whether one glyph corresponded to more than one phoneme. It was for eliminating ambiguity (2.1 (3) of this article). The relation between all the single consonants and vowels and the glyphs is almost uniquely identified; however, it is not easy to identify the relation between a phoneme and a figure when more than one character are connected because the original letterform is changed. Therefore, we verified the uniqueness of the relation between the phonemes and the glyphs by the method of assigning all the glyphs to a two-dimensional table of 32 x 32 indicating the combinations of 32 phonemes in the row and column (8 vowels and 24 consonants for each of horizontal writing and vertical writing). As a result, we verified that there was no overlap in the glyphs, and thus all the relations between the glyphs and the phonemes were uniquely identified.

Third, with regard to easy identification of character classes (2.1 (7) of this article), there is no need for verification because vowels, consonants, and others are allocated to definite blocks in the code table.

Fourth, with regard to the consideration of the order of array, as nothing can be done to know the order of array at that time because there is no dictionary at the times when the Middle Uyghur script was used, we cannot verify the matter at this time. Thus, as described above, we can ensure that the code table and the glyphs are good for actual usage and are in line with the principles in designing a code table.

4.2 Effectiveness of Converting Text into Electronic Form in Terms of Philology

We mentioned in 1.3 of this article that the encoding of the Uyghur script would make it possible to analyze the contents of Uyghur literatures quantitatively, and we verified this possibility using the two texts of Abhidharmakosabhasyatika Tattvartha archived

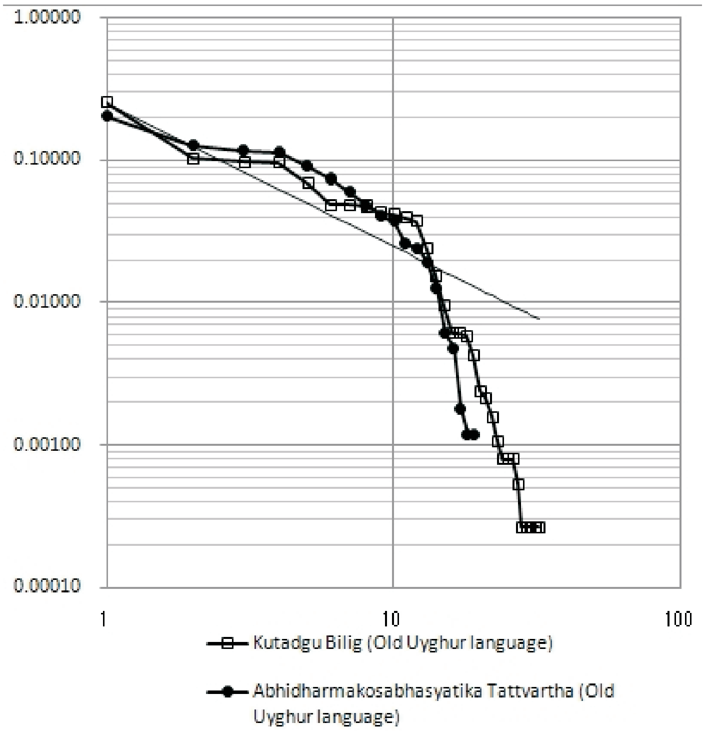


Figure 7: Occurence probability distribution of characters.

in the British Library. The reference numbers of those texts in the library are OR. 8212-75A and OR. 8212-75B, but we refer to them as Text A and Text B according to Masahiro Shogaito mentioned later. We conducted the following verification by inputting the first 6 pages of both the texts. Masahiro Shogaito, who is the authority on the Old Uyghur language and read Abhidharmakosabhasyatika Tattvartha used in this study as a material, points out the ambiguous use cases of characters between the two texts of Abhidharmakosabhasyatika Tattvartha, saying that "For example, though the use of letters for t and d are mixed up in both the texts, yindam[唯] is always spelt yynd 'm while it is always spelt yynt 'm with a few exceptions" [2]. The information about whether such ambiguous differences in the forms of characters occur randomly or systematically may give important clues to the study of philology. We examined whether the ambiguousness found in some words is systematic or not by comparing the appearance frequency in the first 6 pages between Text A and B (Table 8).

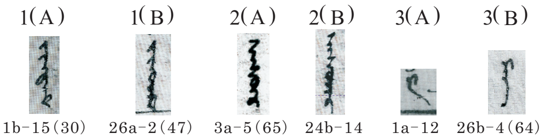
As a result, while the letterforms used for the words of example 1 and 2 are definitely separated between the texts, it is hard to distinct the letterforms used for the word of the example 3 [私] clearly between the texts. This result shows that analysis of computerized text by the use of character codes would make it easy to extract differences in and characteristics of the utilization pattern among texts.

5 Conclusion

In this article, we designed the character code and glyphs for Middle Uyghur script based on the Kutadgu Bilig, which is typical literature written in horizontal Uyghur script, and the Uyghur transcript of Abhidharmakosabhasyatika Tattvartha, which is written in vertical Uyghur script, for the purpose of processing the Middle Uyghur texts electronically. We then conducted verification experiments by inputting the 12 pages in total from Kutadgu Bilig and Abhidharmakosabhasyatika Tattvartha using the horizontal and vertical glyphs of Uyghur script in order to confirm that the Uyghur character code table and glyphs are good for practical use and meet the principles of code design. The result

Table 8: Ambiguous use-case between different texts.

Comparison word			Meaning	British Library collection of books transcript テクストA Or 8212-75A (1a 1-15～3b 76-90) (41p ～ 46p)	 テキストB Or 8212-75B (24b 1-15～27a 76-90) (340p ～ 345p)
1	(A)	Yindam	Nothing but	1	
	(B)	Yintam	Nothing but		1
2	(A)	Sastr	Theory	12	
	(B)	Sastir	Theory		1
3	(A)	Mn	I	2	4
	(B)	Man	I		10



showed that the Uyghur character code and the glyph set were good enough for practical use and they would provide an effective tool for the study on historical literatures. We have sent this Uyghur character code table to the ISO ' s experts of encoding for the review of consistency with the design concept of the international character code table now.

We think this article may provide a model for experts who try to encode scripts in the same situation as the Uyghur script now in terms of describing the basic points to be considered and design procedure in encoding historical scripts whose experts are not so many.

6 Acknowledgments

We would like to express our gratitude for Deborah W. Anderson and Unicode Technical Committee.

References

[1] Tadahiko Goto(Supervision), Tomoji Taniguchi(Volume), Digital . Archivist outline, Japanese education publication, 2006, Tokyo.

[2] Sinasi Tekin, Sources of Oriental Languages and Literatures, ABHIDHARMA-KOSA-BHASYA-TIKA TATTVARTHA-NAMA, Garland Publishing, Inc. New York, N.Y, 1970.

[3] Masahiro Shogaito, Studies in the Uyghur Version of the Abhidharmakosabhasya-tika Tattvartha, Volume1, Volume2, Volume3, Shokado, 1993.

[4] Yoshiki Mikami, A History of Character Codes in Asia, Joint Publication, 2002, Printed in Japan.

[5] Kobayashi Tatsuo, Koichi Yasuoka, Satoshi Tomura, Yoshiki Mikami, (編著), Character-code of days of the Internet, Joint Publication, 2001, Printed in Japan.

[6] The Association for Natural Language Processing (NLP)), Encyclopedia of Natural Language Processing, Printed in Japan, 2009.

[7] http://www.unicode.org/roadmaps/smp/

[8] P. Daniels, W. Bright, The World's Writing Systems, New York, Oxford University Press 1996, Printed in the United States of America.

[9] TURK DIL KURUMU, (I. N. Dilman), Kutadgu bilig Tipkibasim, Viyana Nushasi, Alaeddin Kiral Basimevi, Istanbul, 1942.

[10] G. R. Rachmati, Turkische Turfan-Texte VII, BERLIN 1936.

[11] Takeo Abe, Research History of West Uyghur country, Nakamura print, 1955, Kyoto.

[12] Fumio Yajima(Supervision), Ikko Tanaka(Composition), Man and Writing, 1995, Tokyo.

[13] Character society in the world(Volume), Chart ceremony of character of the world, Yoshikawa Kobun Kan, Tokyo.

[14] Takashi Kamei, Rokuro Kono, Eiichi Chino, The Sanseido Encyclopaedia of Linguistics, Volume 1, 2, 3, 4, Languages of the World, Part Three, First Published 1992, Made and Printed in Japan at the Sanseido Press, Tokyo.

[15] Robert Dankoff, Wisdom of Royal Glory, The University of Chicago Press, Ltd, London, p.1, Published 1983.

[16] Shoyim Bo'tayev, Qutadgu Bilig, Cho'lpon nomidagi nashriyot-matbaa ijodiy uyi, Toshkent, 2007.

[17] A. Dilacar, Kutadgu Bilig Incelemesi, Ankara Universitesi Basimevi-1972.

[18] TURK DIL KURUMU, Kutadgu Bilig Tipkibasim, Fergana Nushasi, Istanbul, Alaeddin Kiral Basimevi, 1943.

[19] TURK DIL KURUMU, Kutadgu Bilig Tipkibasim, Misir Nushasi, Istanbul, Alaeddin Kiral Basimevi, 1943.

[20] Ş.Tekin, Sources of Oriental Languages and Literatures, Garland Publishing, Inc. • New York, 1970.

[21] S.Gorn, R.W.Bemer, J.Green, American Standard Code for Information Interchange, Communications of the ACM, 6(8), 1963.

[22] ISO/IEC TR 15285 An operational model for charac-ters and glyphs, First edition, 1998-12-15.

[23] Michael Everson, Desmond Durkin-Meisterernst, Roozbeh Pournader, and Shervin Afshar, Second revised proposal for encoding the Manichaean script in the SMP of the UCS ISO/IEC JTC1/SC2/WG2 N4029R, L2/11-123R, 2011-05-10.

[24] Omarjan Osman, Proposal for encoding the Uyghur script in the SMP of the UCS, ISO/IEC JTC1/SC2/WG2, 2011-11-07.

[25] Ablahat Ibrahim, Spoken Uyghur, University of Washington Press, Seattle and London, Printed in the United States of America. 1991.

[26] Ildiko Beller-Hann, The Written and the Spoken, Berlin, 2000.

[27] http://www.uyghurcongress.org/en/

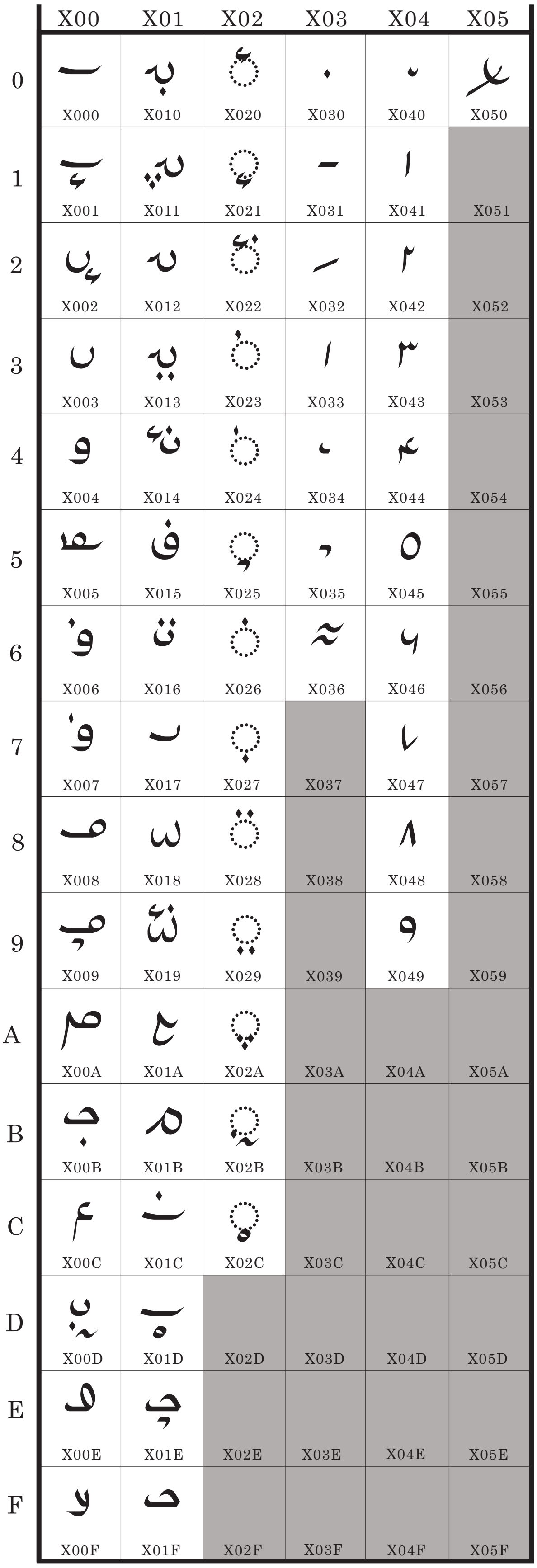
[28] http://uyghuramerican.org/

[29] http://www.rfa.org/english/

Appendix

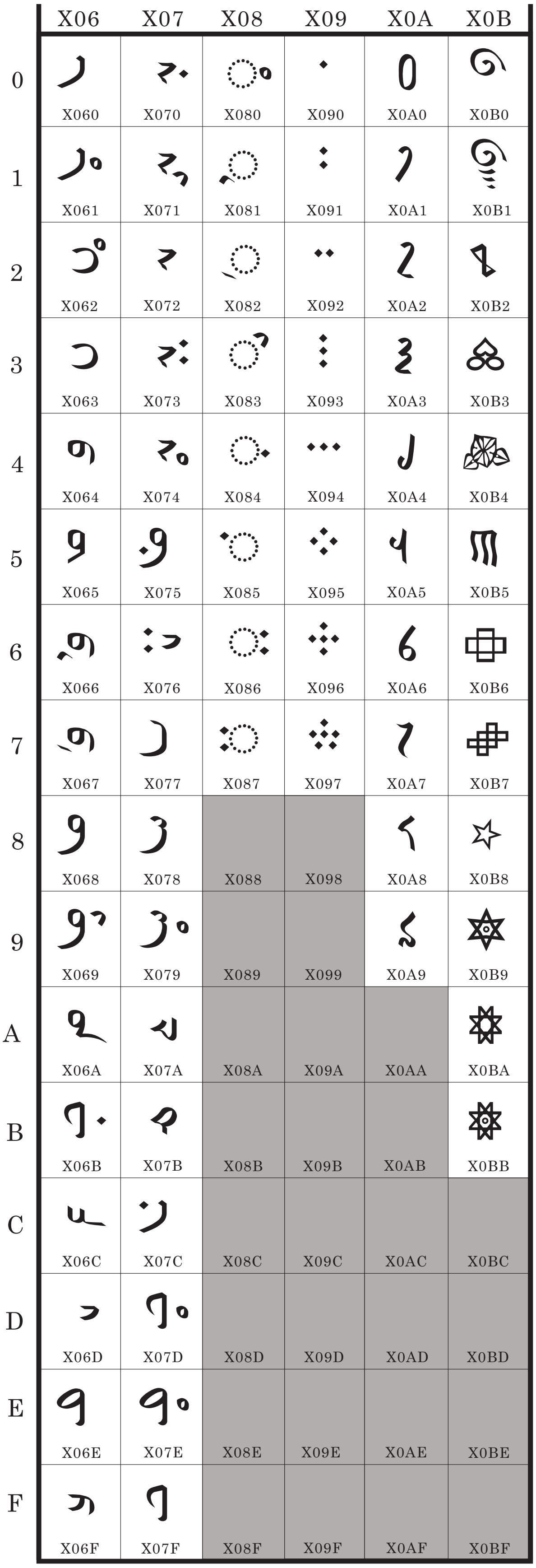
Table 9: Proposed Uyghur Character Horizontally and Vertically Code-Table.

X000-X050 Uyghur



	X00	X01	X02	X03	X04	X05
0	X000	X010	X020	X030	X040	X050
1	X001	X011	X021	X031	X041	X051
2	X002	X012	X022	X032	X042	X052
3	X003	X013	X023	X033	X043	X053
4	X004	X014	X024	X034	X044	X054
5	X005	X015	X025	X035	X045	X055
6	X006	X016	X026	X036	X046	X056
7	X007	X017	X027	X037	X047	X057
8	X008	X018	X028	X038	X048	X058
9	X009	X019	X029	X039	X049	X059
A	X00A	X01A	X02A	X03A	X04A	X05A
B	X00B	X01B	X02B	X03B	X04B	X05B
C	X00C	X01C	X02C	X03C	X04C	X05C
D	X00D	X01D	X02D	X03D	X04D	X05D
E	X00E	X01E	X02E	X03E	X04E	X05E
F	X00F	X01F	X02F	X03F	X04F	X05F

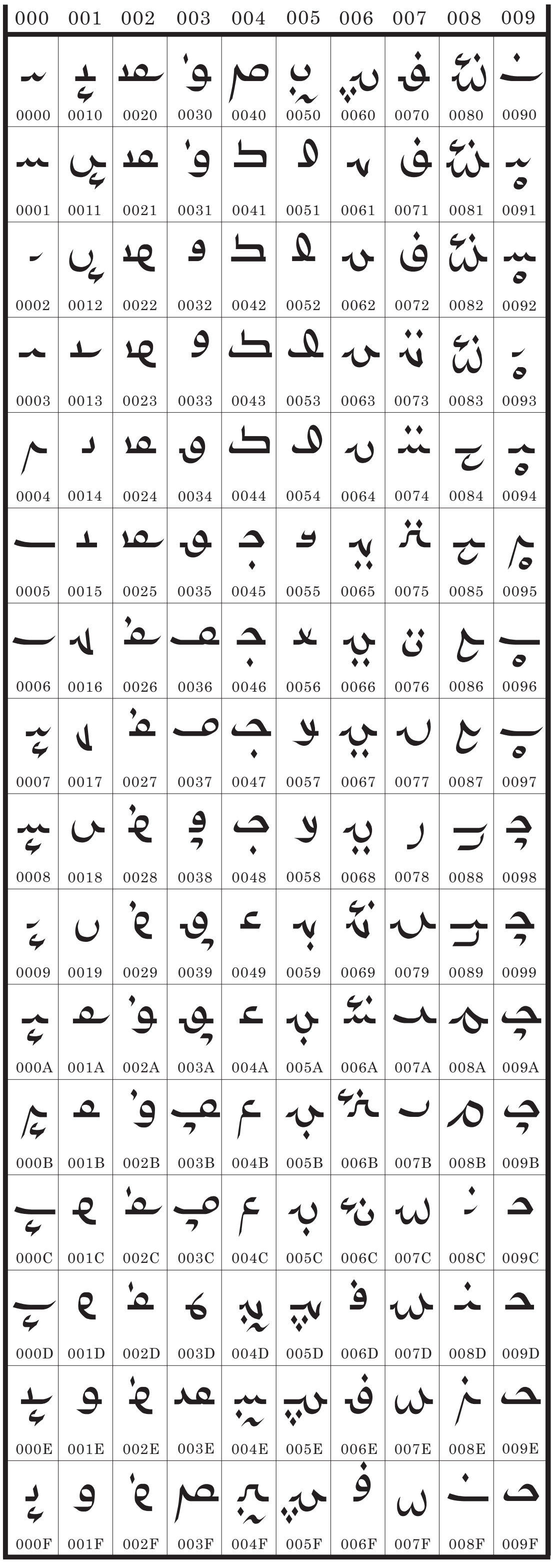
X060-X0BB Uyghur



	X06	X07	X08	X09	X0A	X0B
0	X060	X070	X080	X090	X0A0	X0B0
1	X061	X071	X081	X091	X0A1	X0B1
2	X062	X072	X082	X092	X0A2	X0B2
3	X063	X073	X083	X093	X0A3	X0B3
4	X064	X074	X084	X094	X0A4	X0B4
5	X065	X075	X085	X095	X0A5	X0B5
6	X066	X076	X086	X096	X0A6	X0B6
7	X067	X077	X087	X097	X0A7	X0B7
8	X068	X078	X088	X098	X0A8	X0B8
9	X069	X079	X089	X099	X0A9	X0B9
A	X06A	X07A	X08A	X09A	X0AA	X0BA
B	X06B	X07B	X08B	X09B	X0AB	X0BB
C	X06C	X07C	X08C	X09C	X0AC	X0BC
D	X06D	X07D	X08D	X09D	X0AD	X0BD
E	X06E	X07E	X08E	X09E	X0AE	X0BE
F	X06F	X07F	X08F	X09F	X0AF	X0BF

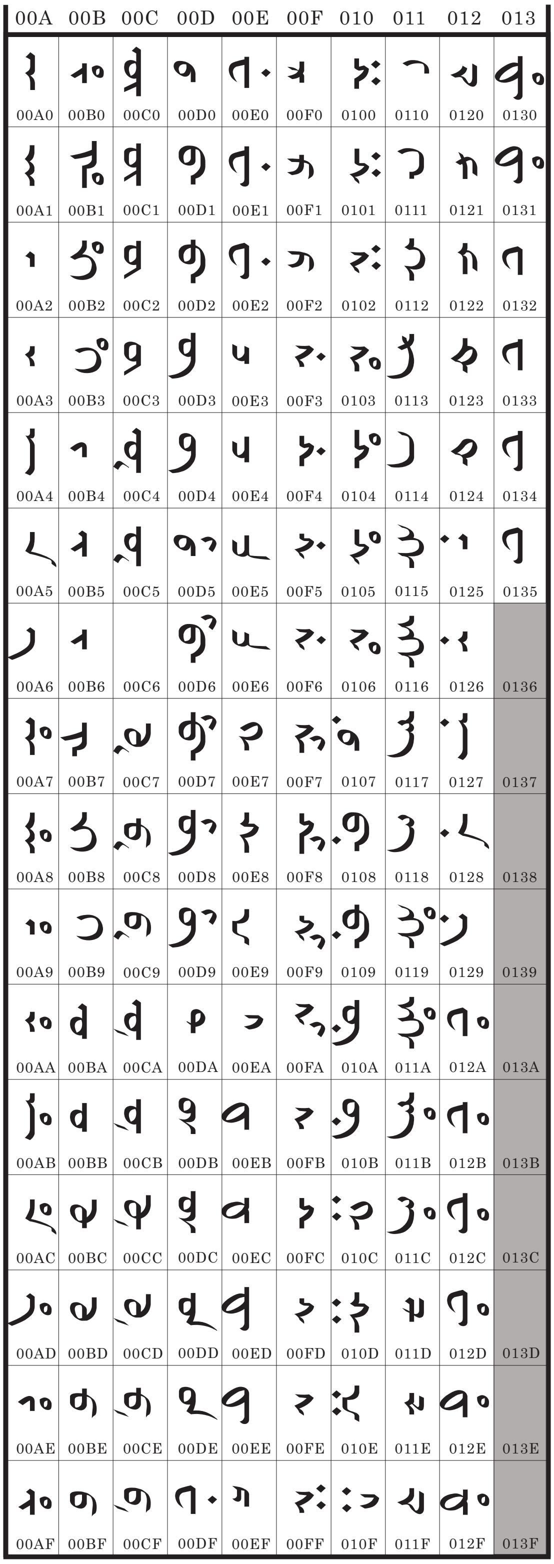
Table 10: Proposed Uyghur Character Horizontally and Vertically Glyph Code-Table.

0000-009F Uyghur



	000	001	002	003	004	005	006	007	008	009
0	0000	0010	0020	0030	0040	0050	0060	0070	0080	0090
1	0001	0011	0021	0031	0041	0051	0061	0071	0081	0091
2	0002	0012	0022	0032	0042	0052	0062	0072	0082	0092
3	0003	0013	0023	0033	0043	0053	0063	0073	0083	0093
4	0004	0014	0024	0034	0044	0054	0064	0074	0084	0094
5	0005	0015	0025	0035	0045	0055	0065	0075	0085	0095
6	0006	0016	0026	0036	0046	0056	0066	0076	0086	0096
7	0007	0017	0027	0037	0047	0057	0067	0077	0087	0097
8	0008	0018	0028	0038	0048	0058	0068	0078	0088	0098
9	0009	0019	0029	0039	0049	0059	0069	0079	0089	0099
A	000A	001A	002A	003A	004A	005A	006A	007A	008A	009A
B	000B	001B	002B	003B	004B	005B	006B	007B	008B	009B
C	000C	001C	002C	003C	004C	005C	006C	007C	008C	009C
D	000D	001D	002D	003D	004D	005D	006D	007D	008D	009D
E	000E	001E	002E	003E	004E	005E	006E	007E	008E	009E
F	000F	001F	002F	003F	004F	005F	006F	007F	008F	009F

00A0-0135 Uyghur



	00A	00B	00C	00D	00E	00F	010	011	012	013
0	00A0	00B0	00C0	00D0	00E0	00F0	0100	0110	0120	0130
1	00A1	00B1	00C1	00D1	00E1	00F1	0101	0111	0121	0131
2	00A2	00B2	00C2	00D2	00E2	00F2	0102	0112	0122	0132
3	00A3	00B3	00C3	00D3	00E3	00F3	0103	0113	0123	0133
4	00A4	00B4	00C4	00D4	00E4	00F4	0104	0114	0124	0134
5	00A5	00B5	00C5	00D5	00E5	00F5	0105	0115	0125	0135
6	00A6	00B6	00C6	00D6	00E6	00F6	0106	0116	0126	0136
7	00A7	00B7	00C7	00D7	00E7	00F7	0107	0117	0127	0137
8	00A8	00B8	00C8	00D8	00E8	00F8	0108	0118	0128	0138
9	00A9	00B9	00C9	00D9	00E9	00F9	0109	0119	0129	0139
A	00AA	00BA	00CA	00DA	00EA	00FA	010A	011A	012A	013A
B	00AB	00BB	00CB	00DB	00EB	00FB	010B	011B	012B	013B
C	00AC	00BC	00CC	00DC	00EC	00FC	010C	011C	012C	013C
D	00AD	00BD	00CD	00DD	00ED	00FD	010D	011D	012D	013D
E	00AE	00BE	00CE	00DE	00EE	00FE	010E	011E	012E	013E
F	00AF	00BF	00CF	00DF	00EF	00FF	010F	011F	012F	013F

Table 11: Part of Kutadgu Bilig, [9].

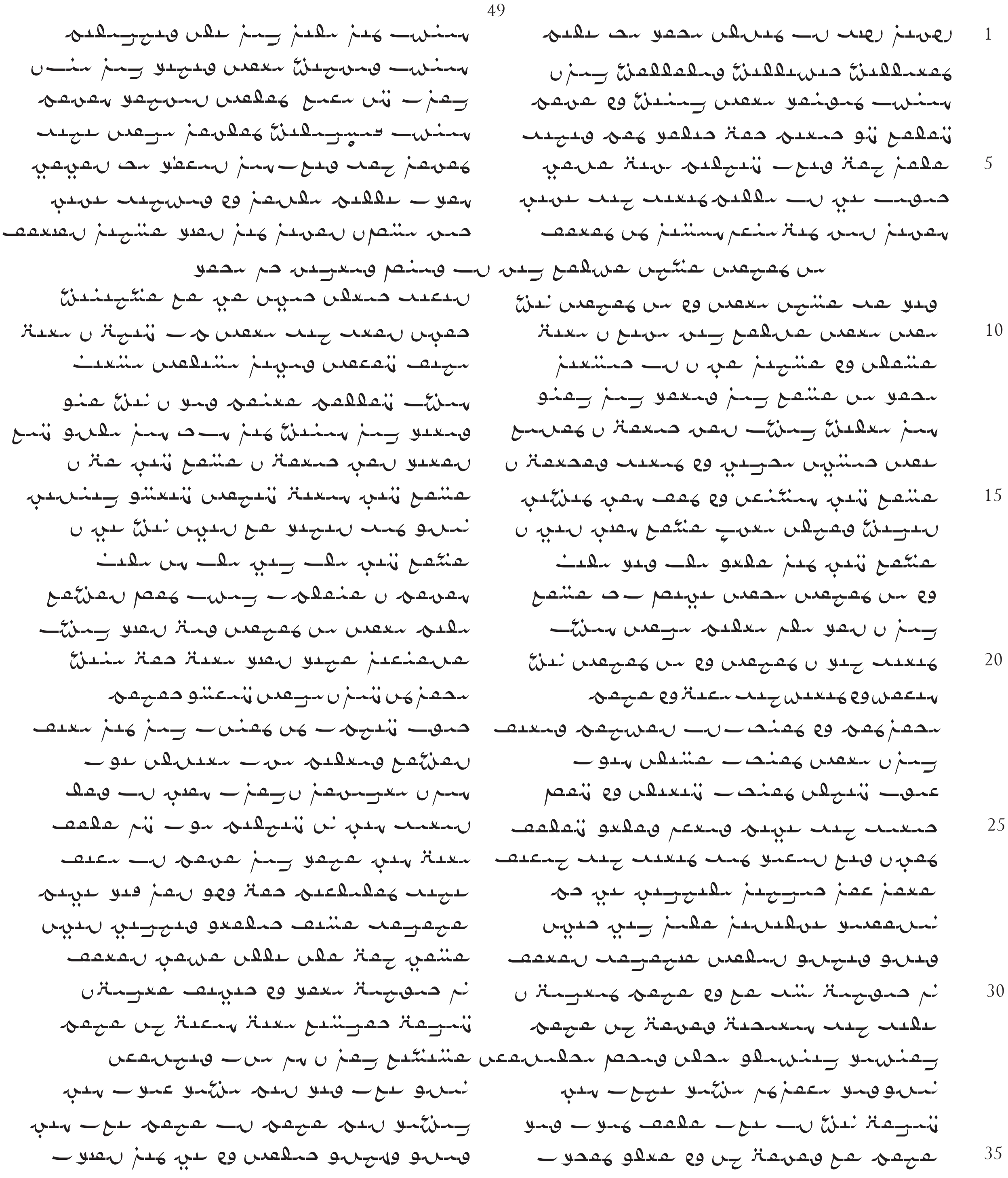


Table 12: Part of Abhidharmakosabhasya-tika Tattvartha, 1a 1-15[2],[3].

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15