L2/20-191
2020-12-18

Final proposal to encode Old Uyghur in Unicode

Anshuman Pandey

pandey@umich.edu
pandey.github.io/unicode

December 18, 2020

Document History

This proposal is a revision of the following:

- L2/18-126: "Preliminary proposal to encode Old Uyghur in Unicode"
- L2/18-333: "Proposal to encode Old Uyghur in Unicode"
- L2/19-016: "Revised proposal to encode Old Uyghur in Unicode"
- L2/20-003R: "Revised proposal to encode Old Uyghur in Unicode"

It incorporates comments made by the UTC Script Ad Hoc Committee and other experts in:

- L2/18-168: "Recommendations to UTC #155 April-May 2018 on Script Proposals"
- L2/18-335: "Comments on the preliminary proposal to encode Old Uyghur in Unicode (L2/18-126)"
- L2/19-047: "Recommendations to UTC #158 January 2019 on Script Proposals"
- L2/20-046: "Recommendations to UTC #162 January 2020 on Script Proposals"

The major changes to L2/20-003R are as follows:

- Change of default orientation of the script from vertical to horizontal (§ 4.2)
- Expanded description of letters, signs, and punctuation using specimens of different script styles (§ 5)
- Enumeration of characters not currently proposed for encoding (§ 7)
- Confirmed the right-joining behavior for *zayin*; previous proposals defined it as dual-joining
- Withdrawal of merged letters for handling ambiguity; recommendation to use mark-up instead (§ 8.5)
- Withdrawal of letters for terminal variants (§ 5.4, § 7); additional analysis is required
- Withdrawal of the space-filling terminal (§ 5.5); additional analysis is required
- Withdrawal of signs used for Arabic transliteration (§ 7); to be proposed later with related signs
- Withdrawal of the stem extending character in favor of using the existing Arabic TATWEEL
- Description of additional dot-like and dotted punctuation signs (§ 5.7)
- Description of terminal connections between words and recommendation for representation (§ 5.2)
- Explanation for the allocation of four columns to the proposed script block (§ 6.1)

A previous version of this proposal was reviewed by the following experts:

- Yukiyo Kasai (Centrum für Religionswissenschaftliche Studien, Ruhr-Universität Bochum)
- Dai Matsui (Graduate School of Letters, Osaka University)
- Mehmet Ölmez (Department of Modern Turkic Languages and Literatures, Istanbul University)
- Nicholas Kontovas (Leiden University)

Support for the proposed encoding has been expressed by scholars:

- L2/20-199: "Endorsement of the Old Uyghur encoding proposal L2/20-191" (Kontovas)

1 Introduction

The 'Uyghur' or 'Old Uyghur' script flourished between the 8th and 17th centuries, primarily in the Tarim Basin of Central Asia and throughout various parts of Asia; as far as Anatolia to the west, Mongolia to the east, and Iran and Afghanistan to the south. Originally used for writing medieval Turkic languages, such as Karakhanid (ISO 639-3: xqa) and Chagatai (ISO 639-3: chg), it became a pan-Asian script as its use was expanded for recording languages such as Chinese, Mongolian, Tibetan, and Arabic.

It developed from the 'cursive' style of the Sogdian script during the 8th–9th centuries into an independent writing system with vibrant scribal and print traditions. Styles of the script are classified broadly as 'square' ('formal', 'book') or 'cursive'. Gradations such as 'semi-square' and 'semi-cursive' naturally arose to enlarge the spectrum of styles. Moreover, as usage of the script continued, other styles developed, such as the 'formal' post-Mongolic hand used after the 15th century in Islamic manuscripts. Block printing was developed in the 14th century for producing books of Buddhist texts. Common usage of the script diminished by the 16th century, and was replaced by new orthographies for Turkic languages based upon the Arabic script. However, its usage in Gansu is attested through the 17th century.

The Uyghur script was a medium for textual transmission across linguistic and religious cultures. A vast amount of Uyghur manuscripts are Buddhist texts, but there are also documents with Manichaean, Christian and Islamic content. It was used for recording Turkic literature and for administrative purposes. On account of the culture contacts of its users, it was used alongside other major Asian scripts. There are numerous documents containing the Uyghur script with intralinear Han characters and with interlinear Sanskrit annotations in 'Turkestani' or Central Asian styles of Brahmi. Other biscriptal documents in Old Uyghur contain Phags-pa seals; the Khitan large script; and Arabic, Armenian, and Hebrew scripts are also extant.

Just as Turkic speakers borrowed the script of the Sogdians, other communities in Central Asia borrowed the Uyghur script for writing their languages. As such, Uyghur is situated in the middle of a script continuum that originates from the Sogdian script of the 'Ancient Letters' and terminates at modern Mongolian. A popular narrative regarding the origin of the Mongolian script recounts that the scholar Tata Tonga, a chancellor of the Naiman Khanate, developed an orthography for writing the Mongolian language using the Uyghur script in the 13th century, during the reign of Genghis Khan. The Uyghur-based Mongolian script developed into a distinctive script with its own orthographic conventions, and independent scribal and print cultures.

Western scholars have studied the Uyghur script and its written record since the early 20th century. It was during that time that European expeditions to Turfan unearthed vast amounts of materials in Uyghur and other scripts. German and Russian scholars adapted the Uyghur script for modern typesetting. Texts in the Uyghur script were edited and published by F. W. Max Müller, V. V. Radlov, and others. At least two styles of metal types were produced for printing these editions, based upon the square style used in manuscripts and the style used in block prints. Interest in the Uyghur script has continued to grow steadily, especially during the past two decades with an increase in focus on the cultures, socieites, and polities of and along the Silk Road. Various institutions that obtained materials from Turfan and other sites have digitized their collections or are in the process of doing so, such as the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), British Library, and other institutions associated with the International Dunhuang Project (IDP).

2 Nomenclature

The term ‘Uyghur’ occurs in Old Turkic inscriptions as 𐰆𐰖𐰍𐰆𐰺 *ujğur*; in medieval Turkic documents as 𐽳𐽶𐽲𐽶𐽾 *wyγyr*; ئۇيغۇر *uyğur* and Уйғур in the modern Uyghur language; and 维吾尔 *wéiwú’ěr* in modern Chinese. It has various language-dependent Latin transliterations. It is rendered ‘Ouïgour’ in French and ‘Uigurisch’ in German. There are multiple English spellings, eg. ‘Uighur’, ‘Uigur’, ‘Uygur’, ‘Uyghur’. The *Oxford English Dictionary* and *Merriam-Webster Dictionary* use ‘Uighur’. However, modern scholars who study Central Asia and write in English prefer ‘Uyghur’ (see Mair 2009). This convention aligns with the spelling ‘Uyghur’ recommended by the Terminology Normalization Committee for Ethnic Languages of the Xinjiang Uyghur Autonomous Region (2006).

The term ‘Uyghur script’ applies to both the Sogdian-based script used for medieval Turkic languages and the later Arabic-based orthography used for the modern Uyghur language, which is not directly related to the former languages. The two scripts are distinguished by using the descriptor ‘old’ for the historical script, as a matter of convenience. To be sure, neither ‘Uyghur’ nor ‘Old Uyghur’ is an accurate designation for the script. The renowned Turkologist, Gerard Clauson notes that the “name is probably as anachronistic as that name when applied to the language” (1962: 100). The script had been in use in Central Asia before the Uyghur language became prominent in the 8th century (1962: 43). However, Clauson concludes that “no useful purpose would be served by suggesting some other name” (1962: 100–101).

In this document, ‘Uyghur’ is used as the normative English spelling and the proposed Unicode identifier for the script is ‘Old Uyghur’. The name pertains specifically to the script within the context of Unicode, and it does not refer to any particular language, culture, or community.

3 Encoding History

3.1 Justification for encoding

Although the Old Uyghur script is derived from Sogdian and is the ancestor of Mongolian, and shares similarities with both scripts, it has requirements that justify an independent encoding in Unicode:

- *Distinctive repertoire* The Old Uyghur repertoire has characters that do not exist in Mongolian, such as *zayin*. The names, values, and order of characters do not correspond directly to those of Mongolian, and which reflect Mongolian preferences and pronunciations.

- *Plain text representation* There is a requirement for representing Old Uyghur documents in plain text. Scholars of Central Asia need to distinguish text in Old Uyghur, Sogdian, and Mongolian. Plain-text representation of Old Uyghur should properly convey the distinctive graphical feature of the script.

- *Unification of multiple styles* Old Uyghur has several styles that should be represented using a unified encoding and a single representative style. An independent block for the script provides a means for managing these styles and uniquely representing them in plain text. Unifying Old Uyghur with Sogdian or Mongolian would not provide a means for adequately distinguishing between different styles of these scripts and would lead to ambiguity, especially in plain text where these scripts occur together.

- *Encoding model* The proposed encoding model for Old Uyghur defines characters using palaeographical values, as opposed to Mongolian, which is encoded on a phonetic basis. Moreover, the proposed default orientation for Old Uyghur is horizontal, which enables usage of the script in a common upright, right-to-left orientation, which also differs from Mongolian.

3.2 Previous Unicode proposals

Proposals to encode Old Uyghur were previously submitted to the Unicode Technical Committee (UTC) by Omarjan Osman:

- "Proposal for encoding the Uygur script in the SMP" (L2/12-066)
- "Proposal to Encode the Uyghur Script in ISO/IEC 10646" (L2/13-071)

These proposals provide valuable background on the history and usage of the script, and details about the representation of letterforms and orientations of the script in different manuscripts. Based upon the provenance and attributes of two important sources, Osman identified two major variations of the script along a geographic basis. He describes the 'western' form as being written horizontally from right to left, and an 'eastern' form that is written vertically from top to bottom (p. 11). Osman thought it necessary to accommodate both orientations of the script through character encodinng. Thus, his proposed repertoire contains upright glyphs for the horizontal form and the same glyphs rotated 90 degrees counter-clockwise for the vertical form.

The model presented in L2/13-071 is ambitious, but it is not practical for purposes of character encoding. It is also incompatible with the Unicode character-glyph model. The encoding of separate characters for horizontal and vertical orientations of a letter results in a model that establishes separate semantic values for glyphic variants of a given letter. Such a repertoire is redundant and prone to complications, for example, errors caused by usage of a horizontal letter in a string of vertical characters, etc. It would be more appropriate to consider such glyphs as directional variants instead of separate characters. Moreover, instead of attempting to accommodate orientations of the script at the character level, it would be practical to use mark-up and layout to achieve the desired display. Nonetheless, Osman's proposal is a useful resource for further investigating the requirements for encoding Old Uyghur. His proposed repertoire includes digits and several diacritics (whose exact provenance is not given), which may need to be encoded in order to support the complete representation of Old Uyghur texts.

3.3 Existing standards

There are no existing formal standards for the Old Uyghur script. The closest related digital standard for the script is the Unicode encoding for Mongolian. Recently, the government of China published a standard known as "GB/T 36331-2018 'Information technology – Uigur-Mongolian characters, presentation characters and use rules of controlling characters'". According to Liang Hai, GB/T 36331-2018 is a subset of GB/T 26226-2010, which is China's standard for encoding Mongolian — based upon the complete Unicode encoding for the script — and equivalent to Mongolia's MNS 4932: 2000. Another subset of GB/T 26226-2010 is GB/T 25914-2010, which provides a standard for the modern writing system for the Mongolian language. Given the reference to "Uigur-Mongolian", it is apparent that the standard is intended for the representation of the early stages of the Mongolian script, using the phonemic model of the Unicode encoding and similar glyphs. However, it is not a character-encoding standard for Old Uyghur.

4 Script Details

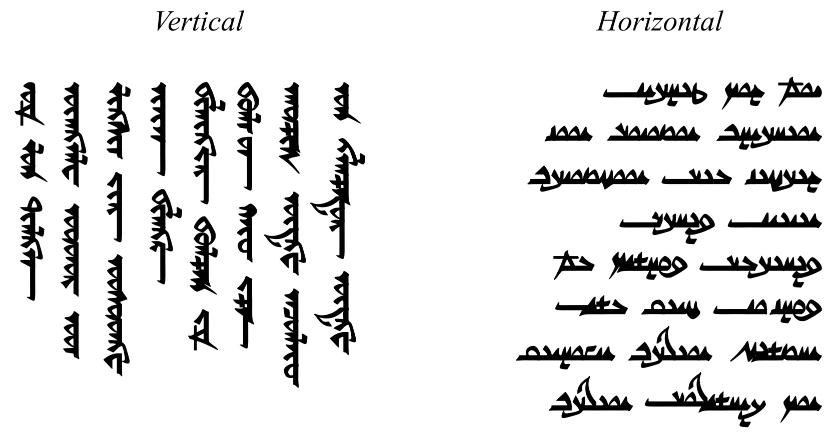
4.1 Structure

The Old Uyghur script is a cursive joining alphabet. The structure is similar to that of Sogdian, with letters joined together at the baseline. The basic letters have an isolated shape and contextual forms when they occur in initial, medial, or final positions. All letters are dual joining, except for *zayin*, which does not join to the left. Diacritics are used for diambiguating letters with similar appearances and for indicating phonetic distinctions between such letters (see § 8.2).

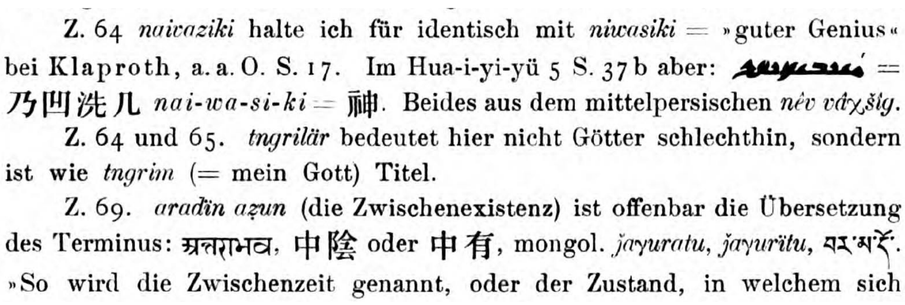
Word boundaries are demarcated using spaces. However, calligraphic space-filling techniques are also used (see § 5.3). Words are generally not broken at line boundaries, nor is there usage of continuation signs. In some texts, a word is split at the end of line and continued on the next line with the next letter in the word. In digital layouts line breaks should occur after words.

4.2 Directionality

The traditional direction of writing for Old Uyghur is vertical, from top to bottom in columns that run from left to right. The vertical orientation is confirmed by biscriptal documents containing Han characters and Central Asian Brahmi. The script is written horizontally in several documents after the 14th century. This may be an influence of the Arabic script. Examples of the script set in both orientations is shown below:



When scholarly printing of Old Uyghur began in the 20th century, some publishers maintained fidelity to the standard vertical orientation (Radlov & Malov 1913), while others used a horizontal orientation for reproductions (Müller 1908). A vertical orientation is practical for blocks of text consisting entirely of Old Uyghur characters. However, a horizontal orientation is convenient for short excerpts of Old Uyghur text, especially when the script occurs in multilingual contexts alongside Arabic, Cyrillic, Devanagari, Tibetan, and other scripts for example, see the excerpt below from Müller (1910: 83):



Z. 64 *naiwaziki* halte ich für identisch mit *niwasiki* = »guter Genius« bei Klaproth, a. a. O. S. 17. Im Hua-i-yi-yü 5 S. 37b aber: = 乃凹洗儿 *nai-wa-si-ki* = 神. Beides aus dem mittelpersischen *nêv vâχšig*.

Z. 64 und 65. *tngrilär* bedeutet hier nicht Götter schlechthin, sondern ist wie *tngrim* (= mein Gott) Titel.

Z. 69. *aradïn azun* (die Zwischenexistenz) ist offenbar die Übersetzung des Terminus: अन्तराभव, 中陰 oder 中有, mongol. *ǰaγuratu*, *ǰaγurïtu*, བར་མ་དོ. »So wird die Zwischenzeit genannt, oder der Zustand, in welchem sich

Given the global range of scholars of Turkic studies and the convenience of representing Old Uyghur text in multilingual contexts, the default orientation for Old Uyghur in Unicode should be horizontal. This is advantageous for representation and display of text in applications that do not support vertical layout.

In the default orientation, Old Uyghur should be oriented horizontally and treated as a right-to-left, top-to-bottom script. Text should be set in horizontal lines that run from right to left, in successive lines from top to bottom. This orientation aligns with the conventional layout for scripts such as Sogdian and Arabic.

The vertical orientation may be handled using layout mechanisms. In vertical mode, the script runs top-to-bottom, in columns that extend left-to-right. Character glyphs would be rotated 90 degrees counter-clockwise.

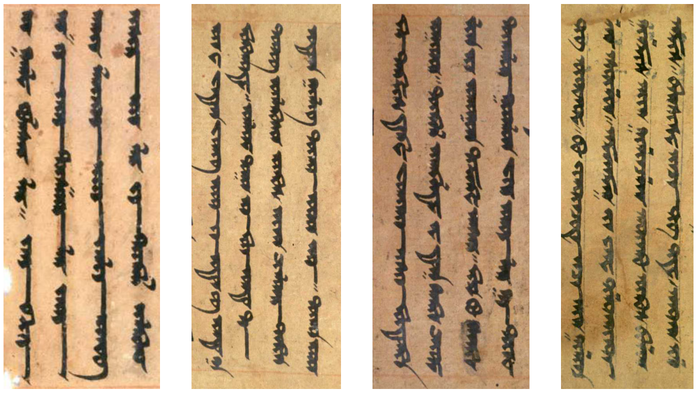
4.3 Styles of the script

Modern scholars classify Old Uyghur documents into two major categories based upon the style of the script: ‘square’ and ‘cursive’ (Moriyasu 2004). There is another style that is not accounted for in this taxonomy, which may be called ‘post-Mongolic’.

The ‘square’ script was used for religious and literary documents from the 9th through 14th century. It was a carefully written formal or book hand, which conveyed the distinctions of letterforms.

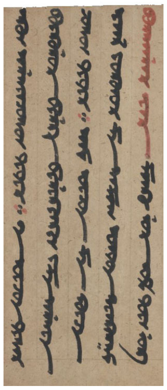
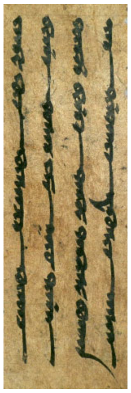
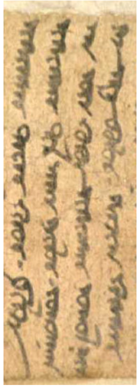
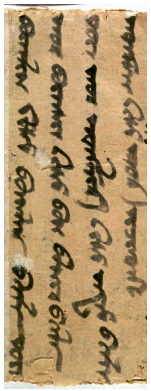
‘square’

Mainz 119 Mainz 841 Mainz 819 U 1071



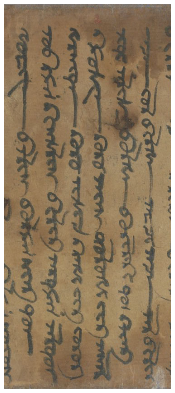
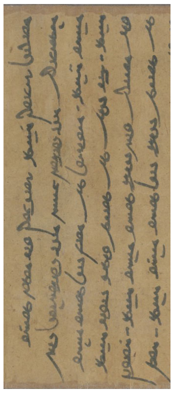
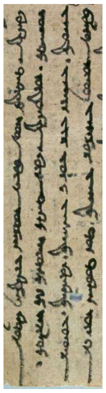
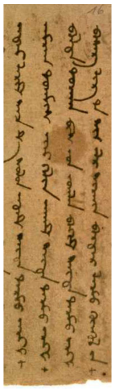
A variant known as ‘semi-square’ is a less formal style of the ‘square’ script.

‘semi-square’

Pelliot ouïgour 13	Mainz 896	U 499	U 560
			

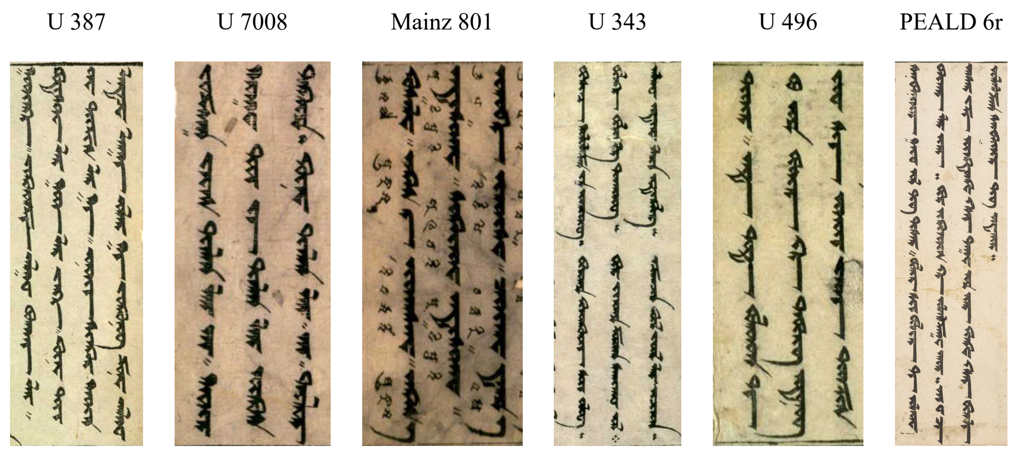
The ‘cursive’ style was used in parallel to the ‘square’ style, and is the running script or general hand for rapid writing. There is less emphasis on the distinctiveness of letters in this style. After the 12th century, this style became the common hand for writing civil documents.

‘cursive’

Pelliot chinois 2998	Pelliot chinois 3046	U 456	U 558
			

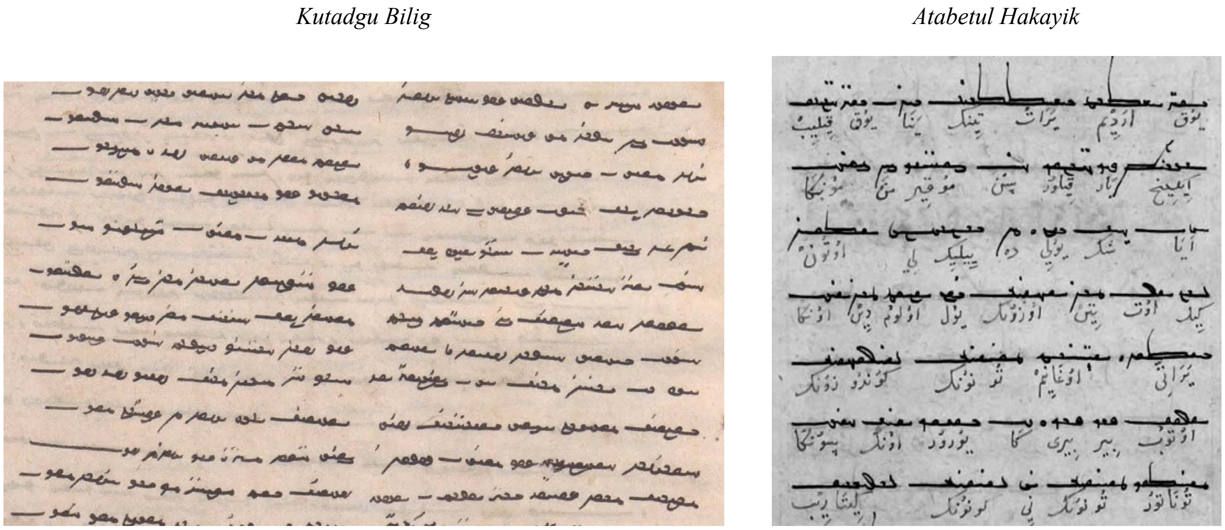
In the 14th century, the Old Uyghur script was adapted for block-printing. The ‘square’ script was used as the basis for the ‘print standard’. This ‘standard’ block-print style is similar to the late inscriptional type, which appears on the stone walls of the Cloud Platform at Juyong Guan, Beijing, erected in the 14th century (see fig. 11). Numerous folios and fragments of block-printed documents have been preserved.

block-print



New styles emerged from the usage of the Old Uyghur script in Afghanistan and Anatolia after the 14th century. This style is a ‘post-Mongolic’ formal hand that was used for literary and civil documents. It is characterized by its miniscule ductus and horizontal orientation:

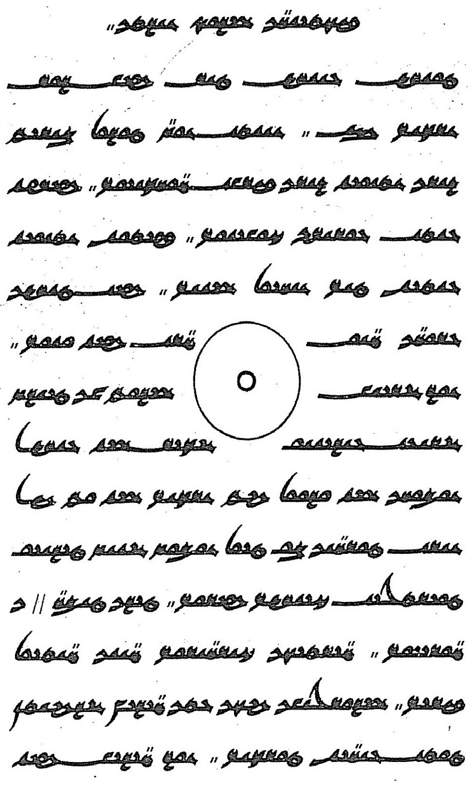
‘post-Mongolic’



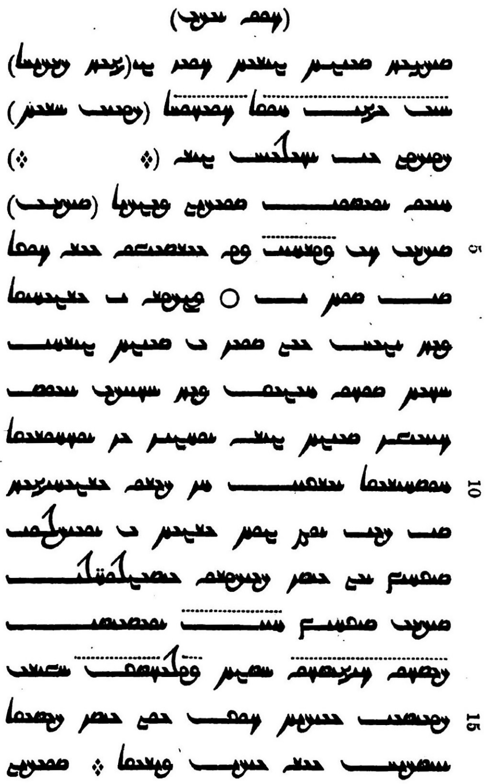
German and Russian scholars adapted the Uyghur script for modern typesetting. Texts in the Uyghur script were edited and published by F. W. Max Müller, V. V. Radlov, and others. At least two styles of metal types were produced for printing these editions, based upon the square style used in manuscripts and the style used in block prints.

European typesetting

Müller (1908)

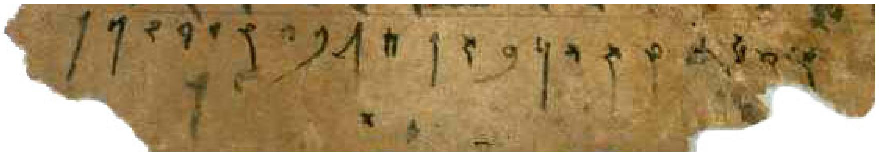


Radlov & Malov (1913)



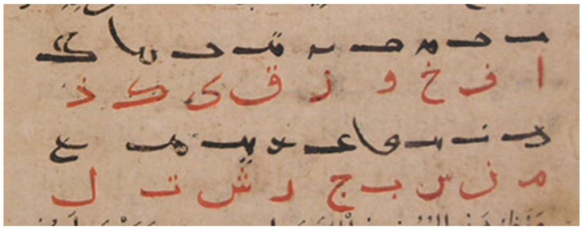
5 Traditional Character Repertoire

The traditional Old Uyghur alphabet consists of 18 letters. There are 15 consonant letters and three that are used for expressing vowels (see § 8.1). The historical repertoire is attested in the manuscript U 40 (see fig. 1), dated to the 9th century:



The inventory contains 21 characters (as read from left to right). The first 17 are basic letters of the script. Following the scholarly nomenclature, these are *aleph*, *beth*, *gimel*, *waw*, *zayin*, *heth*, *yodh*, *kaph*, *lamedh*, *mem*, *nun*, *samekh*, *pe*, *sadhe*, *resh*, *shin*, *taw*. The four letters that follow are not clear due to blemishes in the manuscript. Clauson (1962: 107) suggests that they are 'hooked *resh*', a final *samekh* (or *shin*), a final *mem*, and a two-dotted *heth*.[1] The inventory is important in that it provides attestation for the full repertoire and order of the alphabet, and evidence for the isolated forms of letters, and special forms, eg. final *mem*, two-dotted *heth*. It also provides evidence for the usage of diacritics to expand the alphabet and specify phonetic distinctions, eg. two-dotted *heth* represents /x/ or /q/.

The script of the 11th century is attested in the ديوان لغات الترك *Dīwān luġāt al-turk* "Compendium of the languages of the Turks", a description in Arabic of Turkic languages compiled by the Karakhanid scholar Maḥmūd al-Kāšġarī, in c.1072 (see fig. 2). An excerpt from the text shows Old Uyghur letters (black ink) with their Arabic analogues (red ink):



The repertoire is *aleph*, *beth*, *gimel*, *waw*, *zayin*, two-dotted *heth*, *yodh*, *kaph*, *lamedh*, *mem*, dotted *nun*, *shin*, *pe*, *sadhe*, *resh*, two-dotted *shin*, *taw*, 'hooked r'. The inventory is significant because it indicates the merger of some letters and the usage of additional diacritics for disambiguating such merged forms. Loss of distinctiveness is observed for *samekh* and *shin*, which are represented using a single letter: *samekh* is written using the palaeographical *shin*; *shin* is written using diacritics. The shape of *nun* differs from *aleph* in that it has a less pronounced initial stroke, but the shapes of the letters are close enough that *nun* is denoted using a diacritic. On the other hand, the isolated forms of *gimel* and *heth* are distinctive, but *heth* is written using diacritics; likely for disambiguation in initial and medial positions. Apart from illustrating the dynamic orthography of the script, the attestation is noteworthy because the Arabic transliteration provides a sense

[1] The final *mem* is likely included because it differs in shape from the isolated form; the dotted *heth* has a high frequency of usage. I am not satisfied with Clauson's identification of letters #18 and #19. He states that #18 is the 'hooked' *resh*. While, this letter follows *taw* in the alphabetic order, its shape here resembles 𐽰 — an alternate final form of *aleph* and *nun* that differs from the regular finals — not the 𐾁 'hooked' *resh*. Secondly, he states that #19 is a "final *samekh* (or shin)"; however, these letters do not have a 'special' final shape that differs greatly from their regular finals. I propose that #19 is actually a poorly written 'hooked' *resh*, as supported by the below-base horizontal stroke in the letter.

of the phonetic values of Uyghur letters during this time period in the Karakhanid Khanate. It also indicates that the Uyghur script may have been written horizontally during this period.

The above repertoires are significant for palaeographical reasons in that they show transformations of the script not only over time, but in different regions across Central Asia. Based upon Clauson (1969: 109–110)[2] and details provided by Dai Matsui (personal communication, August 2018–January 2019), the major orthographic practices observed in documents are as follows:

Documents from the 9th century show:

- palaeographic shapes of all 18 letters are distinguishable in good manuscripts
- final *aleph* and *nun* may be written similarly
- initial and medial *gimel* and *heth* are indistinguishable
- two dots above *heth* for representing /q/ or /x/

By the 11th century, the following are observed in some documents:

- *samekh* and *shin* are written in some documents using a single form, resembling *shin*
- two dots beneath *samekh* or *shin* for distinguishing /š/ and /s/
- medial and final *aleph* and *nun* become difficult to distinguish
- in less carefully written documents, final *zayin* may resemble a *nun* without a dot

In addition to the above, other observations in documents of the 14th century include:

- only *kaph*, *lamedh*, *mem*, *pe*, 'hooked' *resh* remain clearly distinctive
- *beth* and *yodh* may not be clearly differentiated or are written using a similar form
- *sadhe* may not be clearly differentiated from *beth* / *yodh*
- *gimel* / *heth* may be indistinct from consecutive *aleph* and/or *nun* without usage of diacritics
- medial and final *taw* indistinguishable from the sequence *waw*-*nun* unless the *nun* is dotted
- *samekh* / *shin* difficult to distinguish from *gimel* / *heth* without dots
- *resh* may be written similarily to consecutive *aleph* and/or *nun*

The above phenomena do not suggest a linear or systematic evolution of the script from the 9th to 14th century. The observations are not uniform across all documents from a given a period or those belonging to a particular style. Rather, variations in orthography may be related to regional scribal practices; the language used by scribes; familiarity of the scribe with the source text being copied, and the accuracy of the source; the type of document being written; and the style of script and degree of careful writing. As shown in Hamilton (2005), there are varying degrees of fidelity to letterforms in 'cursive' documents from the same century. It is difficult to ascertain if the writing of two letters with similar graphical structures using a single ambiguous sign is due to rapid writing; simplification of the repertoire due to assimilation of sounds, for example, loss of sibliants in the languages of scribes, resulting in the merger of *shin* and *samekh* and *shin* due to loss of sibilants; or to formal orthographic reform.

Block-printed documents affirm the non-linear changes to the script. Developed in the 14th century, block-printed Old Uyghur styles are based on the 'square' script, but their repertoires and letterforms are dependently upon on the type-cutter's familiarity with the script. By virtue of being 'printed', such documents

[2] Clauson writes: "In good early manuscripts it is reasonably easy to tell all the eighteen letters apart. Samech and schin have slightly different outlines; initial, and even medial, aleph and nun are just distinguishable, and gimel-cheth, although the two letters themselves are indistinguishable, is identified by two superscribed dots when it represents velar k (or x?)."

imply a ‘standard’ form. However, this form reflects a crystalized repertoire and orthography, which do not account for all palaeographically distinct letters, which occur in earlier documents.

European scholarly printing of Old Uyghur advanced the printing traditions for the script, and also introduced new ways of analyzing the character repertoire. Müller’s *Uigurica* (1908) contains printed reproductions of Turkic literature based upon the original manuscript, in which letterforms were carefully distinguished. This attention to an ‘authentic’ reproduction resulted in Old Uyghur texts without ambigious representations of letterforms. While these may be scholarly texts, they are attestations of Old Uyghur documents and the complete repertoires present in these printed texts may be considered a ‘print standard’ in their own right.

The various charts of the script that have been published in scholarly material provide some assistance in understanding the full repertoire, but some do not fully capture the picture. Of these, Zieme’s chart shows an overview of the representations of letters in different periods (see fig. 7). Other charts, unfortunately, do not provide a full repertoire of attested letters, but appear to be snapshots of the script from a particular document or a period. For instance, von Gabain’s chart shows letters that are typical of the square style (see fig. 5), while Kara’s chart shows letters that resemble those used in block prints (see fig. 9). However, neither of these charts depict all palaeographically attested letters.

The Old Uyghur repertoire underwent simultaneous changes across script styles, regions, and time. But, careful examination of the available sources allows for a complete understanding of the full repertoire and contextual forms of letters.

5.1 Letters

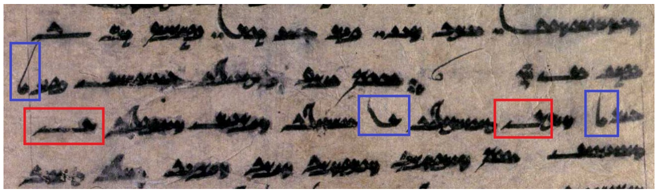
Images in this section have been rotated 90° counter-clockwise for layout purposes.

5.1.1 *aleph* and *nun*

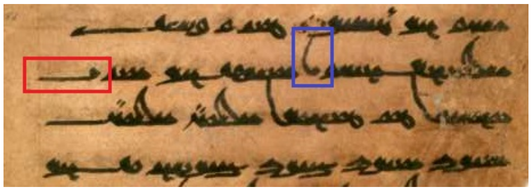
The 𐽰 *aleph* and 𐽺 *nun* are distinctive letters of the script, as attested in U 40 and by Kāšġarī. They are derived, respectively, from Sogdian 𐼰 *aleph* and 𐼻 *nun*. Palaeographically, the body of the Uyghur *aleph* is triangular and has a sharp point at the top; while the Uyghur *nun* is rounded. These two letters present some challenges for character encoding. In some texts their shapes are contrasted in all positions; in others, the distinctions between them are less evident in some positions. It is significant to note that the contrast between these letters is maintained in the printed editions of Uyghur manuscripts in Müller’s *Uigurica* (1908). A description of the letters in various positions is given below:

- *Isolated* A distinctive, isolated *aleph* is a common occurrence and is represented as the regular 𐽰 or the alternate form 𐽰 with a curved terminal. In some cases, the two are used concurrently for distinguishing between final *a* (𐽰) and *ä* or *e* (𐽰) (see the charts in fig. 3, 5). The 𐽰 is not used for *nun*. The excerpts below show regular isolated (red) and the variant (blue) forms:

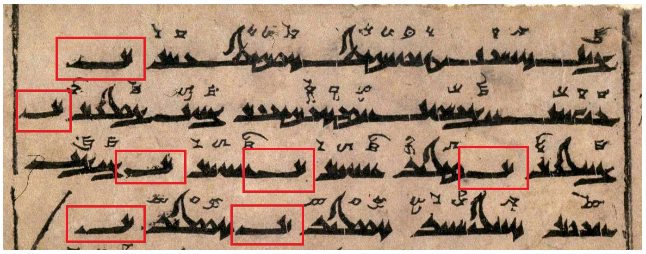
U 3167



Mainz 72

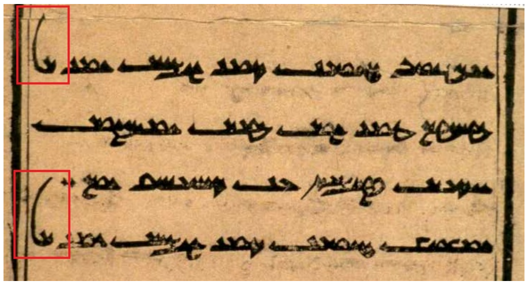


In block prints, the isolated 𐽰 *aleph* is commonly represented as a 'toothed' form 𐽰. This form likely results from creative interpretation of manuscript forms by producers of block prints, but it is a preference that is observed in numerous block prints. This form has a slight resemblance to 𐽷 *kaph*, but is distinguishable by both its shape and context. This 'toothed' form is not used for *nun*. An example of the isolated 'toothed' form is shown below in an excerpt from U 4636:

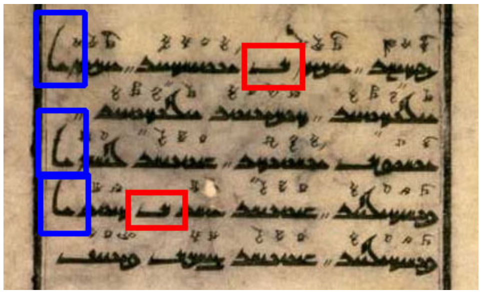


In some block prints, the alternate 𐽰 has a 'toothed' analogue 𐽰, but in others it retains its original shape, even when 𐽰 is used:

U 4708

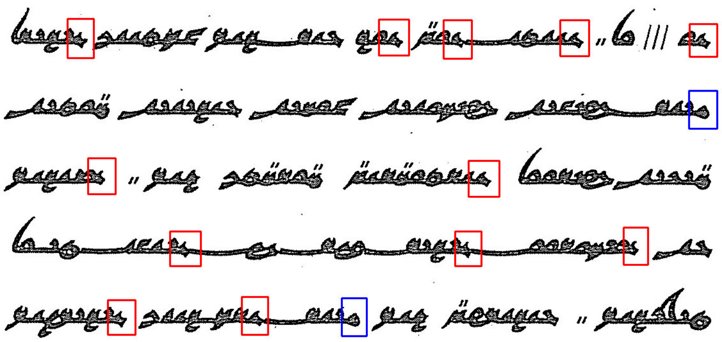


Mainz 801



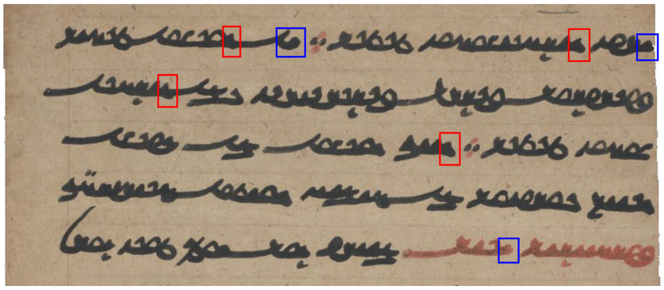
Although there may have been a scribal preference for using 𐽰 instead of 𐽰 in certain contexts, or for choosing the 'toothed' form 𐽰 when the regular angular form 𐽰 is used in the same document, these are considered glyphic variants of 𐽰 at present.

- *Initial* Distinctive forms of initial 𐽰 *aleph* (red) and 𐽺 *nun* (blue) in Müller (1908):

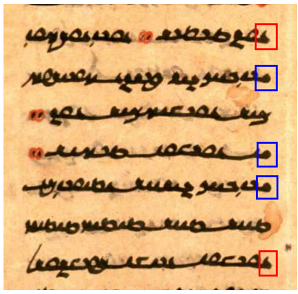


Contrastive representations in semi-square documents of initial *aleph* (red) and *nun* (blue):

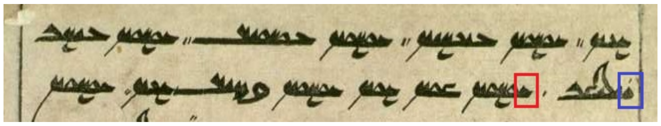
Pelliot Ouïgour 13



Mainz 126

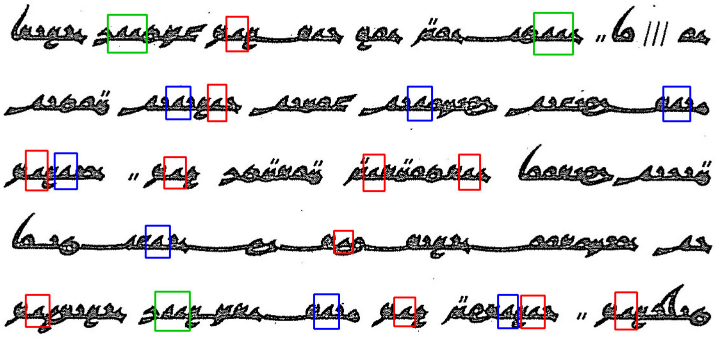


Contrastive representation of *aleph* (red) and *nun* in the block-printed text U 388:



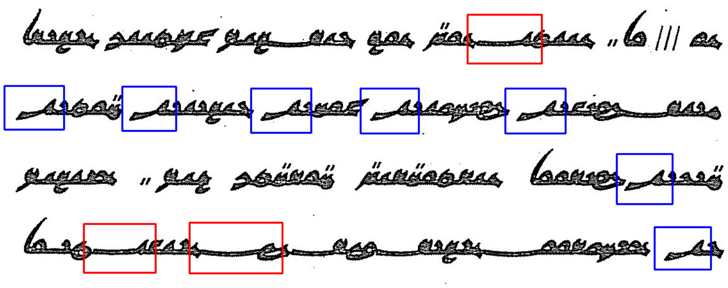
In other documents where contrast between the letters is not well maintained, the initial form of *aleph* may resemble that of *nun*; or initial *nun* may resemble *aleph*; or the two may be written using a generic shape that approximates their structures, such as ▲.

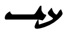
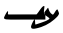
- *Medial* In Müller (1908), there is a clear distinction between the medial 𐽰 *aleph* and medial 𐽺 *nun*, where the former is more hooked and shorter than the latter. The excerpt below shows contrasts between medial *aleph* (red) and *nun* (blue), and sequences of the two letters (green):

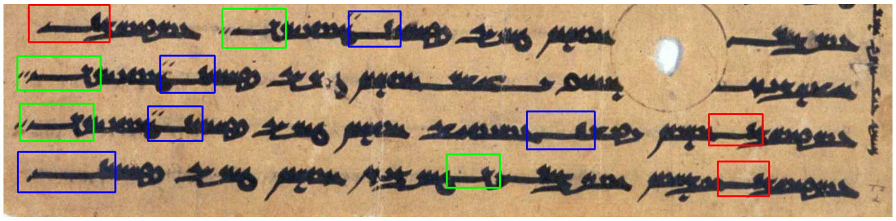


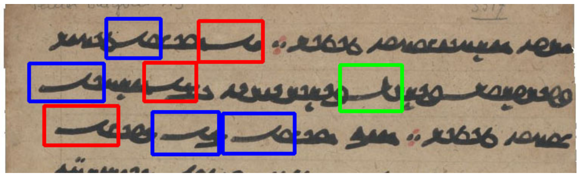
However, in the majority of documents the medial forms are not contrasted. Some perceived lack of contrast may be ascribed to the thick strokes that are characteristic of some scribal practice. Some actual lack of contrast may be due to the ambiguities inherent in cursive or rapid writing where there is less consideration for producing letters carefully. In such cases the medial form of both letters is written using a shape resembling that of *aleph* or *nun*, or a generic shape such as ▲.

- *Final* In Müller (1908), there is a clear distinction between the final 𐽰 *aleph* (red) and final 𐽺 *nun* (blue), where the body of the former is smaller than that of the latter, while the tail of the former is horizontal and that of the latter is curved and slightly curved:

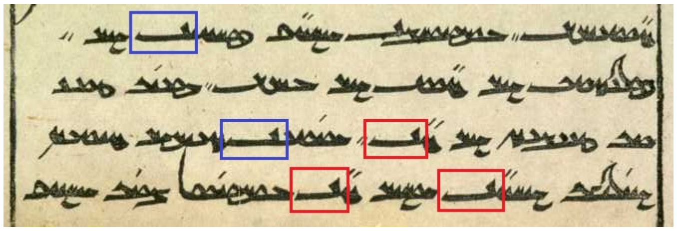


Nonetheless, there are exceptions for representation of final *aleph* following *kaph* or *pe*. In these contexts, *aleph* is represented using its isolated form, eg.  *k* ʾ, instead of the final, eg. *. Even in documents where *aleph* is not distinguished from *nun* in medial or final position, when it follows *kaph* or *pe*, it is written distinctively. Such contrasts are shown below in the excerpts from U 2275 (top) and Pelliot ouïgour 13 (bottom), which show final *aleph* (red) and *nun* (blue), and the distinctive final *aleph* (green) used after penultimate *kaph*:

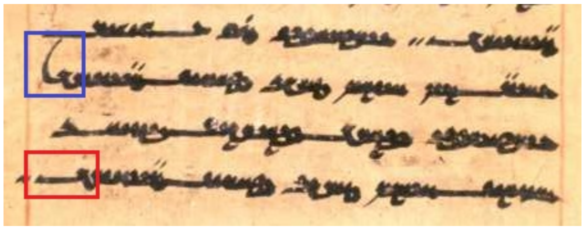




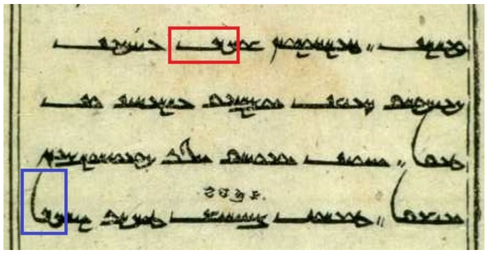
However, in several manuscripts and block prints, the final forms of both letters are written using a single form, such as the below excerpt from U 387:



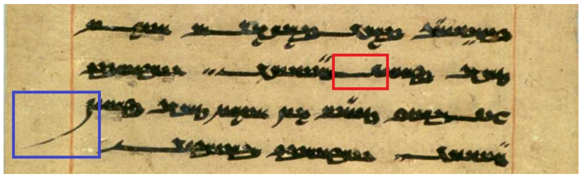
The alternate form is also used with penultimate *kaph*:



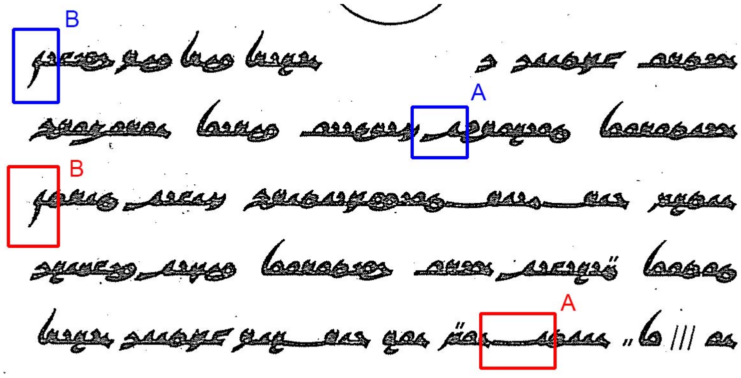
When the 𐽰 / 𐽰 ’toothed’ form of isolated *aleph* is used in a document instead of the regular isolated form 𐽰, it also occurs after penultimate *kaph* and *pe*, as the regular shape of *aleph* (see U 372 below): 𐽷𐽰 / 𐽷𐽰 *k*ʾ, 𐽼𐽰 / 𐽼𐽰 *p*ʾ, compare to 𐽷𐽰 *k*ʾ, 𐽼𐽰 *p*ʾ. Such contextual glyphic variation should be considered conventional behavior.



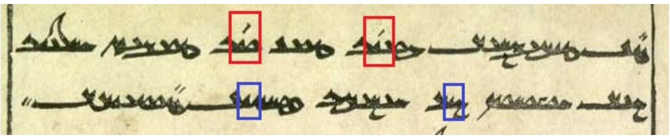
The *aleph* and *nun* is also written as 𐽺 when final. This form occurs concurrently with the regular final *aleph* and *nun* in several manuscripts. It is used at the end of a line or at a text margin when there is limited space for the horizonal terminal of the *aleph* or *nun*. This form may have a semantic function as a morphological separator in certain contexts, but additional research is required in order to make a determination (Matsui, personal correspondence, November 2018). An excerpt from U 947 shows both forms of final *nun*:



As shown in the above discussion of the final forms, *aleph* and *nun* have distinct final forms in Müller (1908). However, the text also shows the two letters written using the same alternate final form 𐽺 (‘B’) at the end of line along with the regular final forms (‘A’):



- *Disambiguation* Due to the ambiguity of these two letters in some documents, the diacritic ◌́ is written above *nun* in order to distinguish it from *aleph* when the two letters are indistinct, as in an excerpt from U 385:



The various forms of *aleph* and *nun* are summarized in the table below:

		X_n	X_f	X_m	X_i
aleph	regular	𐽰	𐽰	𐽰	𐽰
	variants	𐽰 𐽰 𐽰	𐽰 𐽰 𐽰	𐽰	—
nun	regular	𐽺	𐽺	𐽺	𐽺
	variant	—	𐽺 𐽺	𐽺	—

The ambiguity posed by the loss of contrast between *aleph* and *nun* in medial and final positions in various sources adds complexity for uniquely encoding characters that have distinct shapes in some contexts, but that have similar or identical shapes in others. Despite the fact that the rendering of *aleph* and *nun* using a single glyph in various contexts is an inherent aspect of some styles of the writing system, the encoding model should enable a means for uniquely encoding a string containing *aleph* and *nun* such that there is a one-to-one correspondence between a glyph and the identity of the underlying character. Given the above, the following model is practical for encoding *aleph* and *nun*:

		X_n	X_f	X_m	X_i
ALEPH	dual	𐽰	𐽰	𐽰	𐽰
NUN	dual	𐽺	𐽺	𐽺	𐽺

This approach follows the typical model for cursive joining scripts and can distinctively represent all isolated and contextual occurrences of *aleph* and *nun* by encoding them as separate characters on the basis of palaeographical attestations.

- The 𐽰 is to be treated as a glyphic variant of the isolated 𐽰 *aleph*.
- The 'toothed' forms 𐽰 / 𐽰 of *aleph* are to be treated as stylistic variants of 𐽰.
- The final form 𐽰 is to be treated as a glyphic variant of final *aleph* and *nun*, and displayed using a font designed for handling the occurrences of this form.
- The final form 𐽺 used in block prints is to be treated as a glyphic variant of final *aleph* and *nun*, and

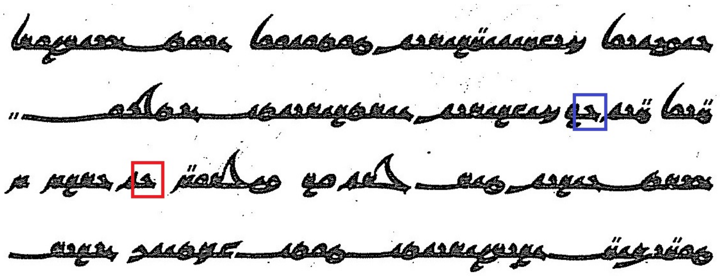
displayed using a font designed for handling the occurrences of this form.

- In block print styles where 𐽰 is used for final *aleph* and *nun*, the form 𐽰 for final *aleph* used after penultimate *kaph* and *pe* should be handled by contextual substitution by the font as part of the regular shaping behavior.

5.1.2 *beth* and *yodh*

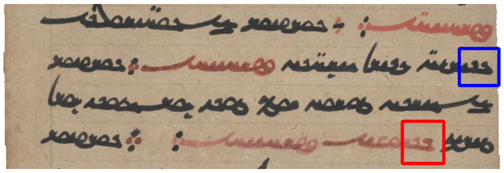
The letters 𐽱 *beth* and 𐽶 *yodh* are palaeographically distinctive letters in the script. They have distinctive forms in all positions, with *beth* possessing either a more angular form and pronounced head, either notched or straight.

Excerpt from Müller (1908) showing distinctive forms of initial *beth* (red) and initial *yodh*:

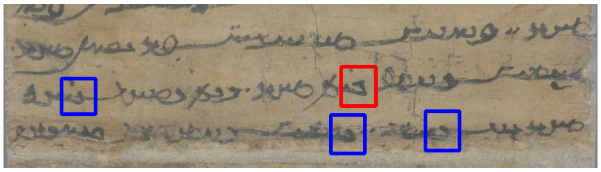


The below excerpts show distinctive forms of initial *beth* (red) and *yodh* (blue) in semi-square (left) and cursive (right) documents:

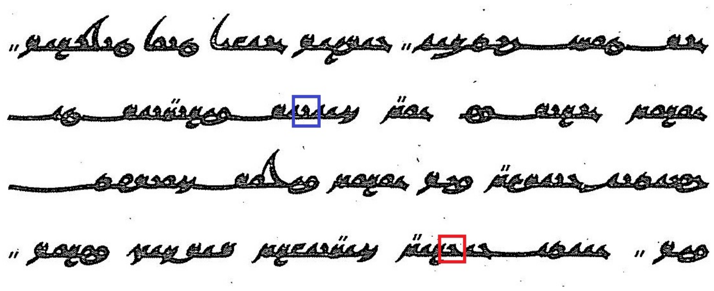
Pelliot ouïgour 13



Pelliot chinois 3049

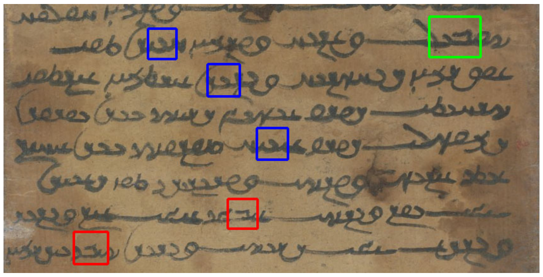


Excerpt from Müller (1908) showing distinctive forms of medial *beth* (red) and medial *yodh*. In this context, the distinctiveness between the letters is more pronounced.

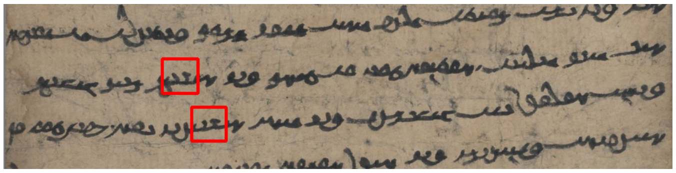


Contrast between medial *beth* and medial *yodh* in sequence in cursive texts:

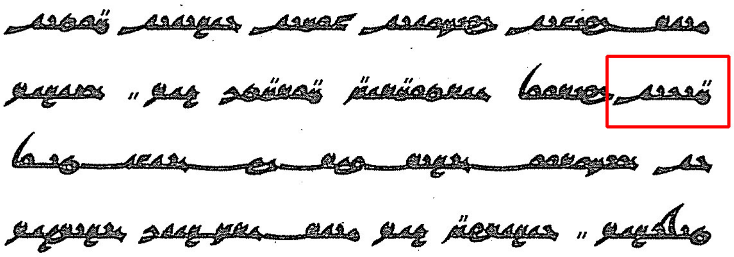
Pelliot chinois 2998



Pelliot ouïgour 3

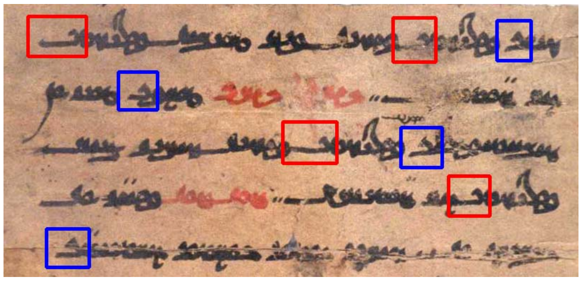


Contrastive representation of a sequence of medial *yodh*, *beth*, *yodh* in Müller 1908 (42, 43). The medial 𐽱 *beth* has a more angular stroke than the medial 𐽶 *yodh*.

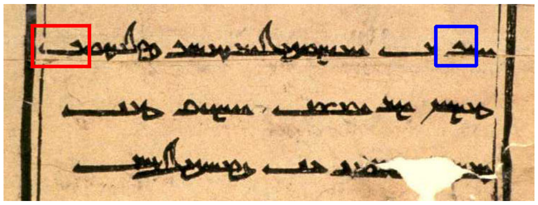


Contrast between final 𐽱 *beth* (red) and final 𐽶 *yodh* (blue) in a block print (from U 4708). The variant final 𐽱 form of 𐽱 *beth* with a left-ward tail, contrasted with final 𐽶 *yodh* in a semi-square document and a block print (U 4708):

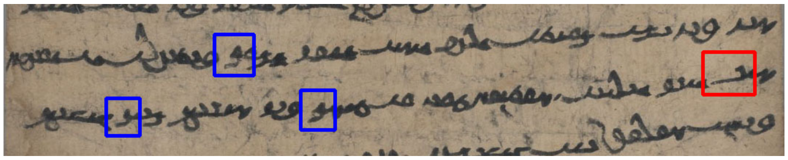
U 5101

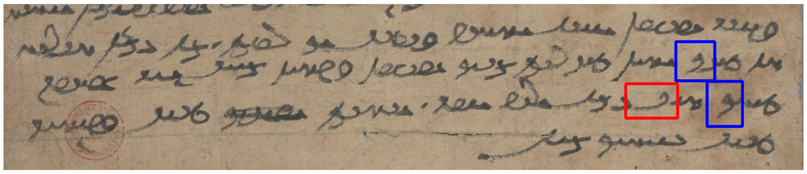


U 4708

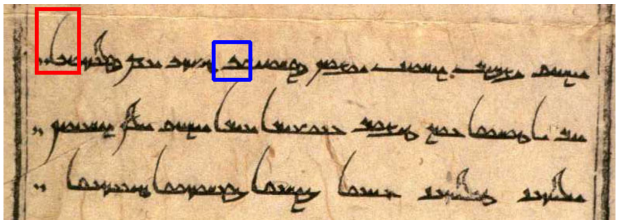


Contrastive representation of final 𐽱 *beth* (red) and final 𐽶 *yodh* (blue) in cursive script, from Pelliot ouïgour 3 (left) and 5 (right). The *beth* is characterized by the length of its terminal, while *yodh* is characterized by both the shape of the body and its short terminal.

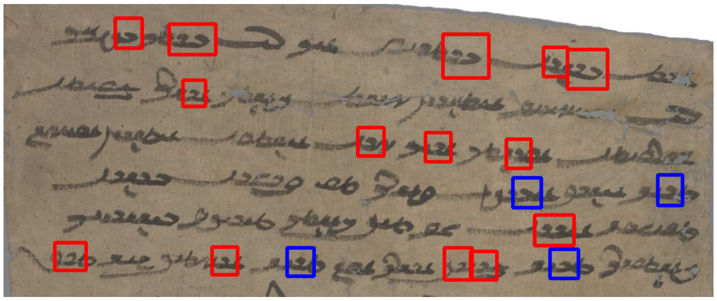




The regular final form of *beth* is ‍𐽱; however, it is also written as ‍𐽱. The curved tail is used likely for distinguishing ‍𐽱 *beth* from ‍𐽶 *yodh* when there is a limitation of space for extending the final stroke of the former. This curved form is to be treated as a stylistic variant.

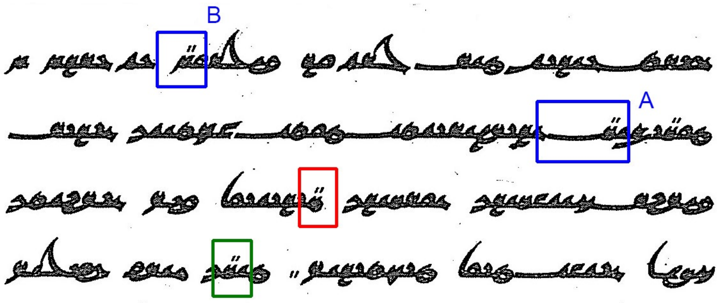


In some less carefully written documents, these two letters are written using an ambiguous form ‍𐽶 that approximates the general outline of the two letters; see the representations of *beth* (blue) and *yodh* (red) in Pelliot ouïgour 2, below. Such cases of ambiguity should be treated as specified in § 8.5.

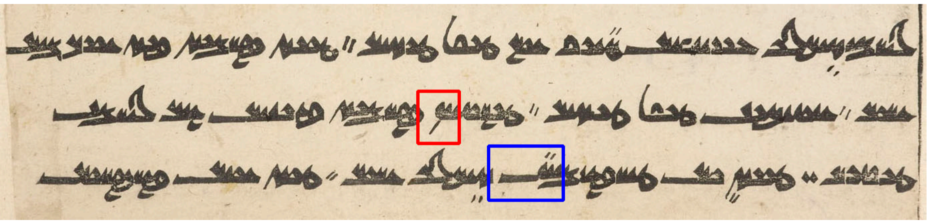


5.1.3 *gimel* and *heth*

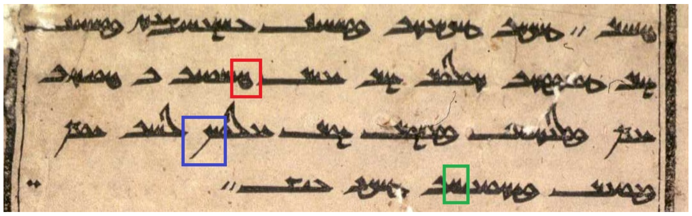
As evidenced in U 40 and Kāšġarī, the letters *gimel* and *heth* are written using the glyphs 𐽲 and 𐽵, respectively. The *gimel* is used for expressing /γ/ and *heth* for /x/ and /q/. Apart from the contrast in isolated and final contexts, the two letters share the same 𐽲‍ initial and ‍𐽲‍ medial shape. In some documents, final *heth* is written using the same final shape ‍𐽲 for *gimel*, but ‍𐽵 is not used for *gimel*.



Generally, *heth* is distinguished from *gimel* using diacritics, eg. 𐽲𐾂, 𐽲𐾄, ‍𐽵𐾂, ‍𐽵𐾄 (see § 8.2 for additional details). These diacritics may be used even when the letter shapes are distinct, such as in PEALD 6a, which shows *gimel* (red) and *heth* (blue):



Initial and medial *gimel* and *heth* have the same form. The excerpt below from U 4680 shows *gimel* as it typically appears in block prints, in initial (red), medial (green), and final (blue) positions:

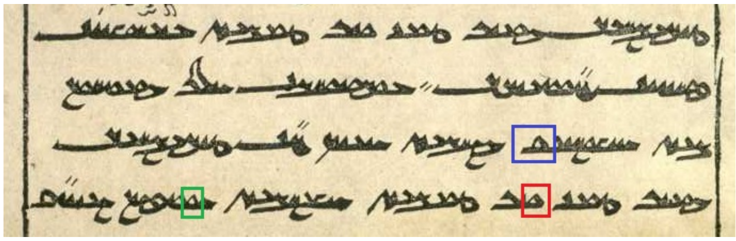


While the initial and medial forms of *gimel* and *heth* are identical, and in some cases, the final form of *heth* is the same as that for *gimel*, the isolated form 𐽵 must be distinguished from 𐽲 in order to represent the distinctive letters of the script, as shown in U 40 and Kāšġarī. To enable the complete representation of these two letters, the following model is proposed:

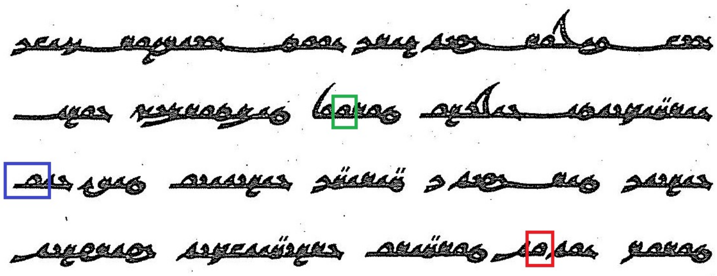
		X_n	X_f	X_m	X_i
GIMEL-HETH	dual	𐽲	‍𐽲	‍𐽲‍	𐽲‍
FINAL HETH	right	𐽵	‍𐽵	—	—

5.1.4 *waw*

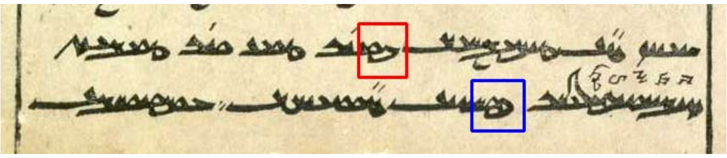
The letter 𐽳 *waw* is consistently represented in Old Uyghur documents. The excerpt below from U 386 shows *waw* as it typically appears in block prints, in initial (red), medial (green), and final (blue) positions:



The following shows the initial (red), medial (green), and final (blue) forms of *waw* used in Müller (1908):



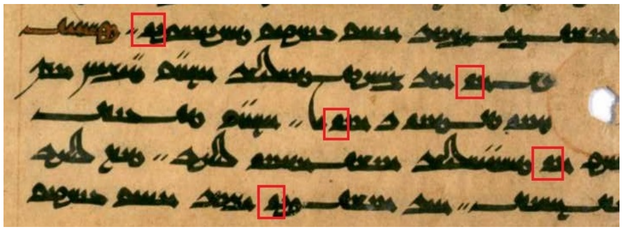
In some calligraphic styles, such as that in the block print U 385, below, when *waw* follows *kaph* (red) and *pe* (blue), the tails of the latter curve into the body of the *waw* to produce a ligature, for instance *kaph* + *waw* may be written as 𐽷𐽳 instead of 𐽷𐽳, and *pe* + *waw* as 𐽼𐽳 instead of 𐽼𐽳.



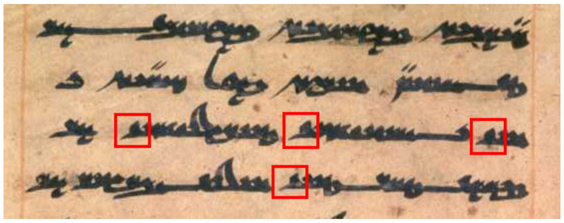
5.1.5 *zayin*

The *zayin* does not join to a following letter. It has the form 𐽴 in the 'square' style:

Mainz 250

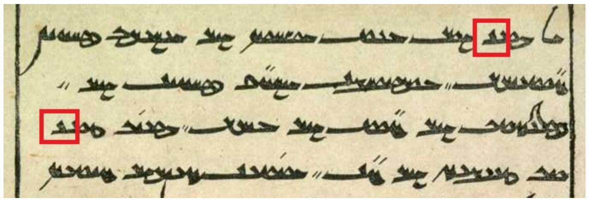


Mainz 119

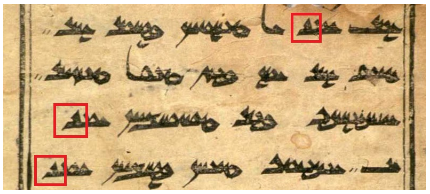


It has the more angular shape 𐽴 in block-printed documents:

U 387

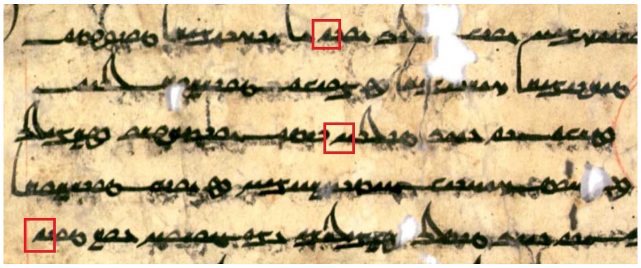


U 4710

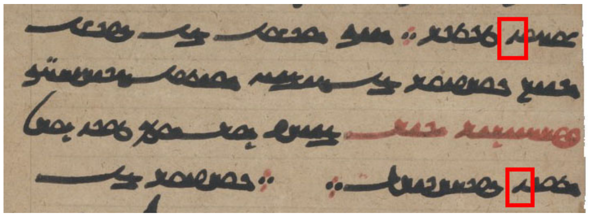


A triangular or 'sawtooth' form 𐽴occurs in semi-square documents:

Mainz 341

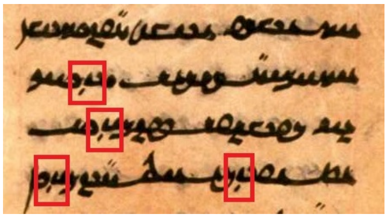


Pelliot Ouïgour 13

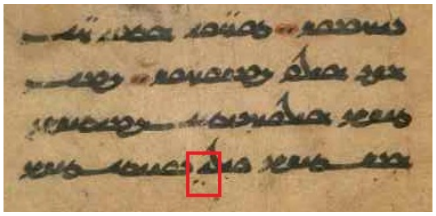


The diacritics ◌𐾃 and ◌𐾅 may be used for indicating /ž/ and other values , eg. 𐽴𐾃 and 𐽴𐾅 (see § 8.2):

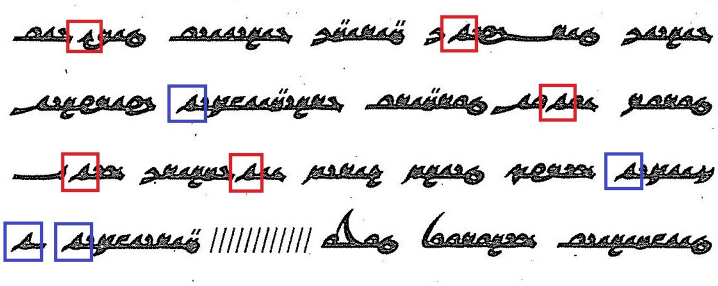
Mainz 126



U 49

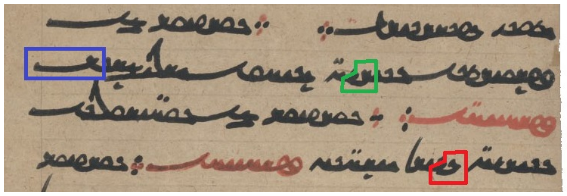


The following shows the form of *zayin* used in Müller (1908). Word-medial (red) and word-final (blue) forms are highlighted specifically to show the right-joining nature of *zayin*:

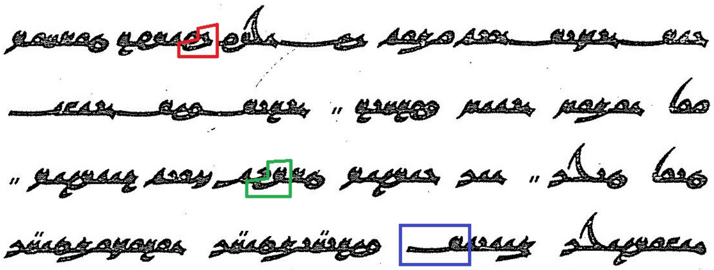


5.1.6 *kaph*

The letter 𐽷 *kaph* has a vertical terminal when isolated and 𐽷 final, but it curves to the right of the baseline and connects below the following letter when 𐽷 initial (red) and medial 𐽷 (blue), as shown in the below excerpt from Pelliot Ouïgour 13:

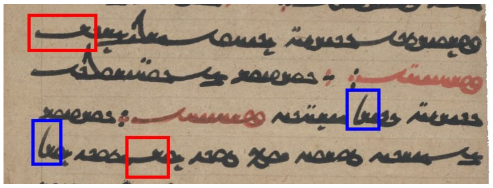


The following shows the initial (red), medial (green), and final (blue) forms of *kaph* used in Müller (1908):

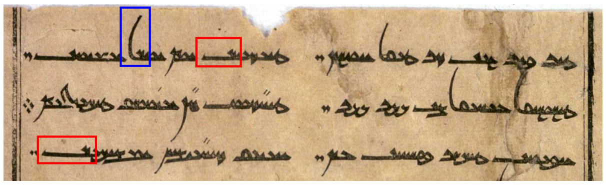


The regular final form of *kaph* is 𐽷, however, the final is also written as 𐽷. The left-ward orientation of the tail is used to accommodate space constraints on a line. It is to be treated as a stylistic variant. Shown below is usage of the regular (red) and the alternate final (blue) in a manuscript and block print:

Pelliot Ouïgour 13

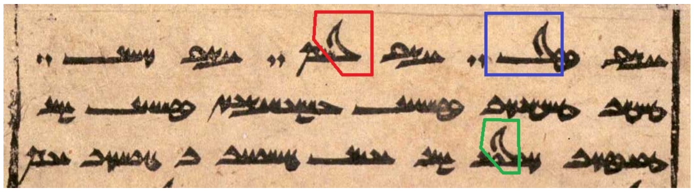


U 4301

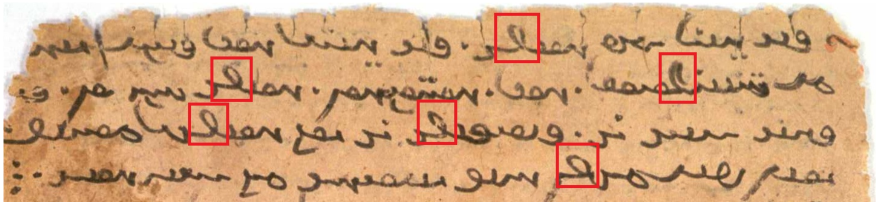


5.1.7 *lamedh*

The letter 𐽸 *lamedh* is consistently represented in Old Uyghur documents. The excerpt below from U 4680 shows *lesh* as it typically sppears in block prints, in initial (red), medial (green), and final (blue) positions:



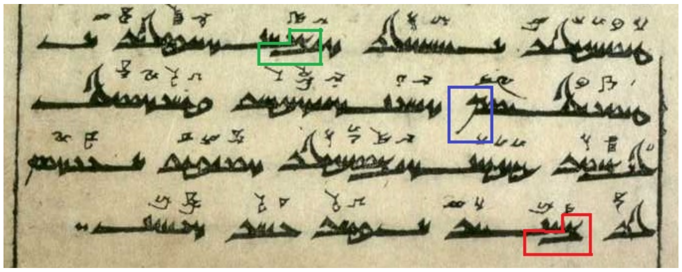
In cursive documents, the top-hook of *lamedh* is curved back towards the baseline and is written as a loop, as in the excerpt from Mainz 91:



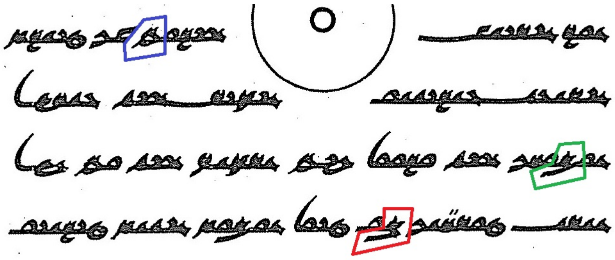
This form is a glyphic variant belonging to the cursive style. The first highlighted *lamedh* in the second line resembles a reversed form of the regular glyph, but it is a poorly-written looped form, and is not semantic distinct from the other instances of the letter.

5.1.8 *mem*

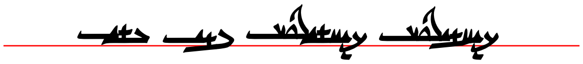
As attested in the inventory in U 40, the *mem* has two distinctive graphemes: 𐽹 and 𐽹. These are the isolated and final forms, respectively. Following the representations in U 40 and Kāšġarī, the 𐽹 has been selected as the isolated form for MEM. Following the cursive joining model, the final form would be rendered when *mem* occurs in final position in a string. The excerpt below from U 351 shows *mem* as it typically sppears in block prints, in initial (red), medial (green), and final (blue) positions:



The following shows the initial (red), medial (green), and final (blue) forms of *mem* used in Müller (1908):



The extender of *mem* extends below the baseline in initial 𐽹 and medial 𐽹 positions. The rendering of preceding and following letters is a stylistic mattter. In the 'square' style, the extender of medial *mem* is written at an angle that slopes downward; the baseline of the preceding letter may also slope such that it joins the extender of *mem*. The following letter connects to the head of *mem*, which may results in all successive letters being written above the baseline, with *mem*, in effect, creating a secondary baseline within a word:

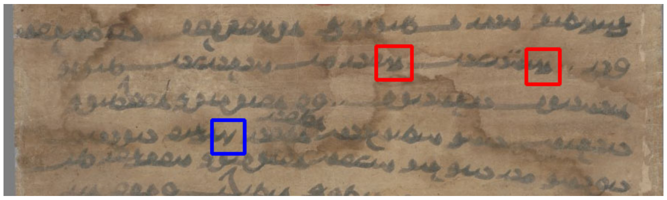


5.1.9 *samekh* and *shin*

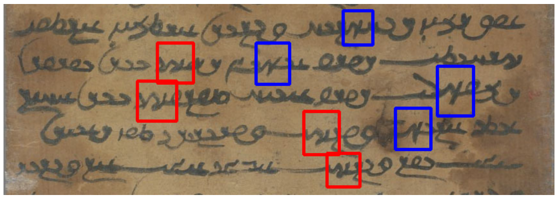
As shown in U 40, the letters 𐽻 *samekh* and 𐽿 *shin* are palaeographically distinctive letters in the script. They are distinguished as follows in initial, medial and final positions in square and cursive documents:

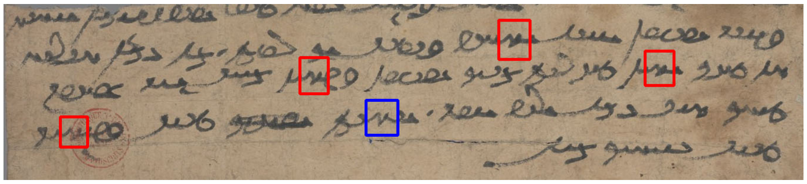
- samekh has a downward stroke that juts down then bows up before curving back down to the baseline
- *shin* has a stroke that angles sharply down then straight back up before descending to the baseline

Distinctive forms of initial *samekh* (red) and *shin* (blue) from a cursive document (Pelliot chinoise 3072):

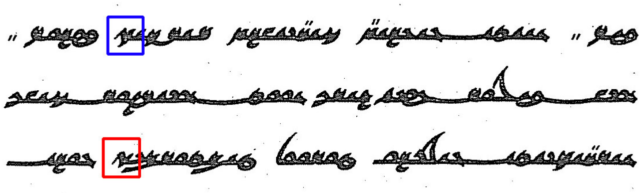


Distinctive forms of medial *samekh* (red) and *shin* (blue) from cursive documents, Pelliot chinois 2998 (top) and Pelliot ouïgour 5 (below):

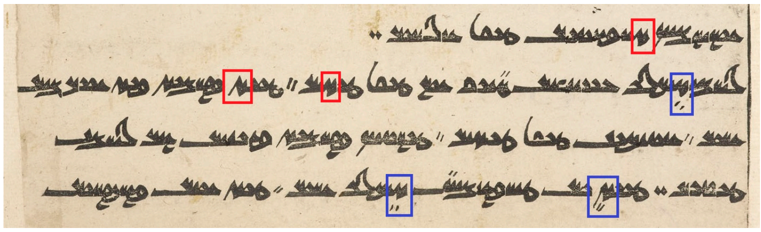




Distinct forms of final *samekh* (red) and *shin* (blue) from Müller (1908):

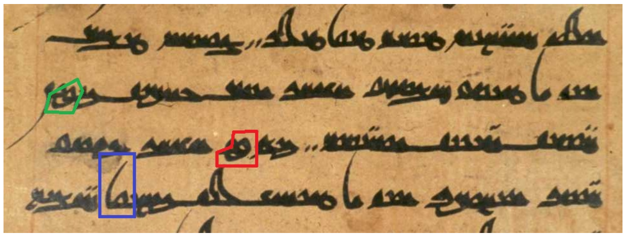


By the 11th century, in some documents, and especially in block-printed texts, both letters were written using a similar glyph based upon the simpler 𐽿 *shin* instead of 𐽻 *samekh*. In such documents, as shown in the excerpt from PEALD 6a, the diacritic ◌𐾅 is applied to 𐽿 *shin* to express /š/, eg. 𐽿𐾅, or ‘marked’ or ‘dotted’ *shin* (see also § 8.2):

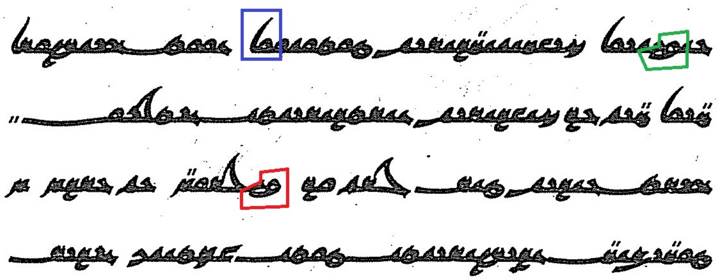


5.1.10 *pe*

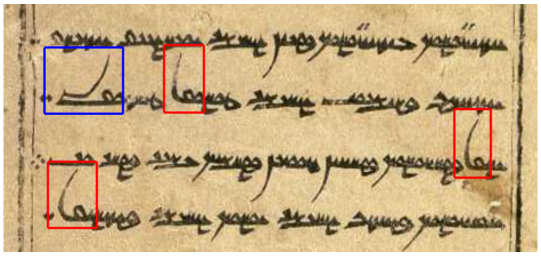
The letter 𐽼 *pe* has a vertical terminal when isolated and 𐽼 final, but its terminal curves up from below the baseline and connects beneath the following letter when 𐽼 initial (red) and medial 𐽼 (blue), as shown in the below excerpt from Mainz 841:



The following shows the initial (red), medial (green), and final (blue) forms of *pe* used in Müller (1908):



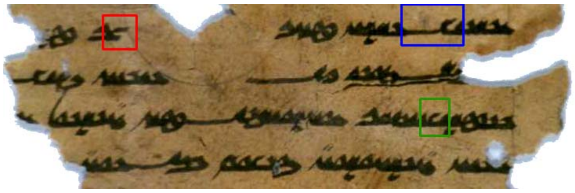
Some manuscript and block-print documents show final *pe* written as 𐽼 (blue) in addition to the regular final form 𐽼 (red), as in an excerpt from U 4162 below:



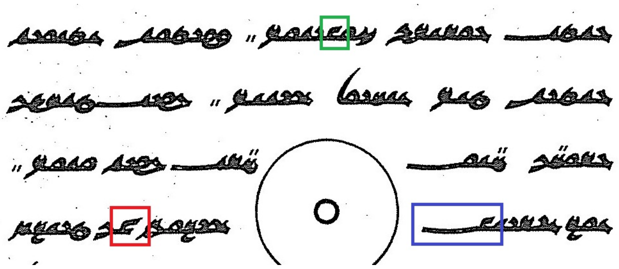
The 𐽼 may be a space-filling terminal; additional research is required to determine its encoded representation (see details in § 5.5).

5.1.11 *sadhe*

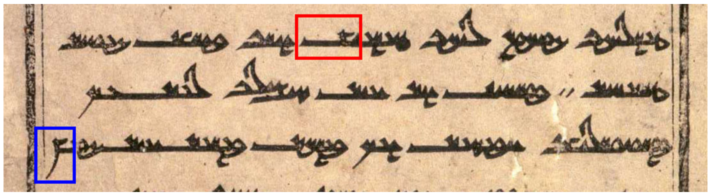
The letter 𐽽 *sadhe* is consistently represented in square, semi-square, and block-print documents. The excerpt below from U 408 shows *sadhe* as it typically appears in square script, in initial (red), medial (green), and final (blue) positions:



The following shows the initial (red), medial (green), and final (blue) forms of *sadhe* in Müller (1908):

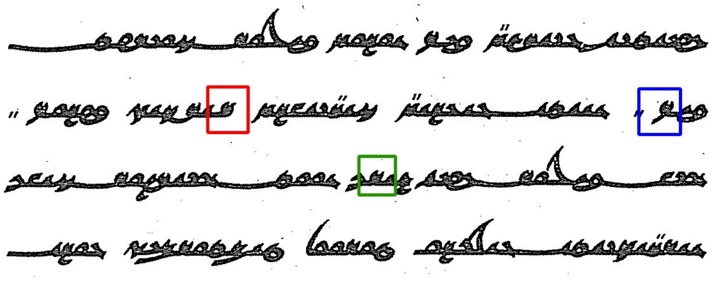


The regular final form of *sadhe* is 𐽽, however, the final is also written as 𐽽, as shown below in an excerpt from U 4680. The left-ward orientation of the tail is used to accommodate space constraints on a line. It is to be treated as a stylistic variant.



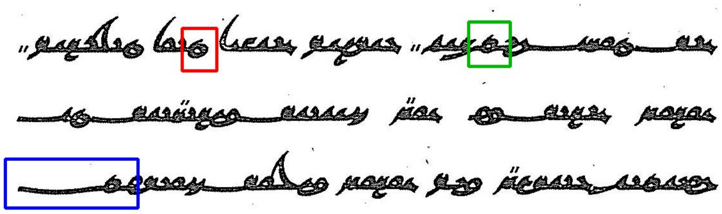
5.1.12 *resh*

The letter 𐽾 *resh* is consistently represented in square script and block-printed documents. The excerpt below from Müller 1908 (44) shows *resh* as it typically appears in initial (red), medial (green), and final (blue) positions in square script:

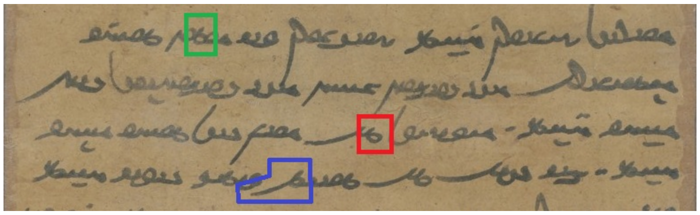


5.1.13 *taw*

The letter 𐾀 *taw* is consistently represented in square script, cursive, and block-printed documents. The excerpt below from Müller 1908 (44) shows *taw* as it typically appears in initial (red), medial (green), and final (blue) positions:



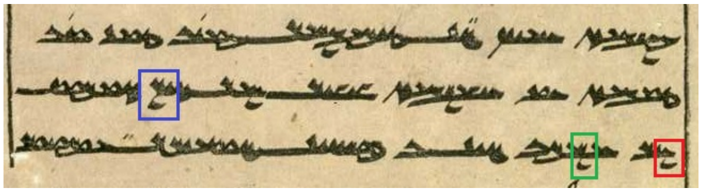
The excerpt below from Pelliott chinoise 386 shows *taw* as it typically appears in cursive documents:



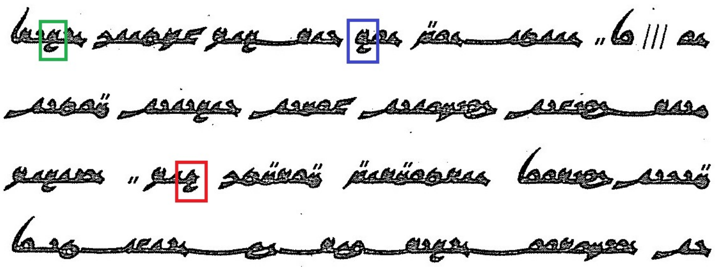
The body of the initial form 𐾀 sits below the baseline, as compared to its medial 𐾀 and final 𐾀 forms. This practice is exhibited in manuscripts and block prints, and may be accepted as normative behavior. The depth of the body of the initial form differs by source. In some cases, the final stroke of the loop meets the stroke of the next letter at the baseline. In other sources, where the terminal looped stroke of *taw* connects with the initial vertical that produces the spine of the letter, the following letter connects to the initial *taw* where the spine of the *taw* meets the baseline.

5.1.14 ‘hooked’ *resh*

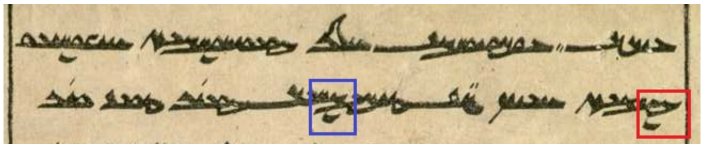
The letter 𐾁 represents the sound /l/. It is derived from 𐽄 U+10F44 SOGDIAN LETTER LESH, which is known as ‘hooked *r*’ (see Pandey 2016b for details). The Uyghur 𐾁 has been assigned the name ‘LESH’, following the name for the corresponding Sogdian letter. This is not a historical name, but one suggested by modern scholars as it aligns with the Aramaic name *resh*, from which it is ultimately derived. The alias ‘hooked *r*’ has been specified in the names list. The excerpt below from U 383 shows *lesh* as it typically sppears in block prints, in initial (red), medial (green), and final (blue) positions:



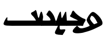
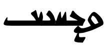
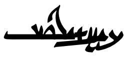
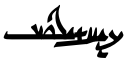
The following shows the initial (red), medial (green), and final (blue) forms of *lesh* used in Müller (1908):



When *lesh* follows *kaph*, *mem*, or *pe*, its hook is attached below the descender of the previous letter. The same excerpt from U 383 shows *lesh* after *kaph* (red) and after *mem* (blue):



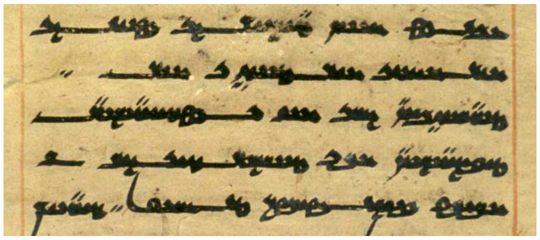
When 𐾁 *lesh* follows letters with elements that extend below the baseline, the hook is detached from *lesh* and placed beneath the extension of the previous letter: 𐽷𐾁 *kaph* + *lesh*, 𐽹𐾁 *mem*, *lesh*, 𐽼𐾁 *pe*, *lesh*. Even if *lesh* does not immediately follow *kaph*, *mem*, or *pe*, its hook may attach to the terminal of the latter for aesthetic considerations, shifted hook vs. static hook, for example:

	static hook	shifted hook
pylyk 'bilig'		
k›lm›dwk 'kälmädük'		

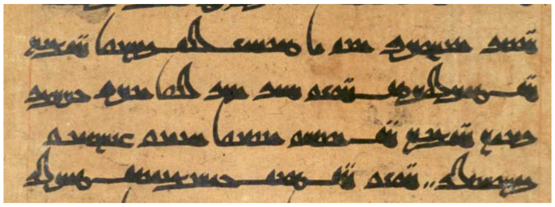
5.2 Terminal extension

In 'square'-script documents, letters with extended horizonal terminals may have their terminals stretched when they occur in word-final position, such that the stroke touches the initial letter of the following word.

U 924

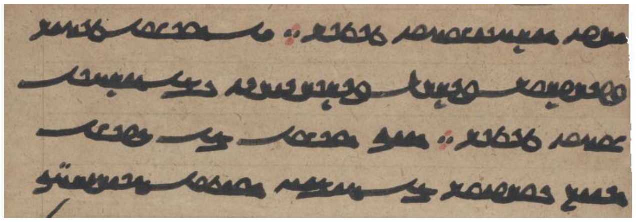


Mainz 841

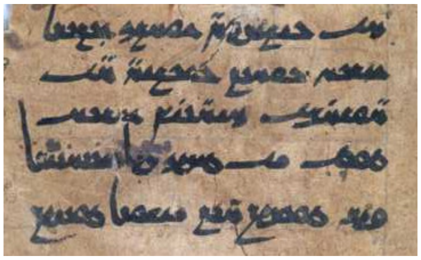


The practice is observed in 'semi-square' documents, such as Pelliot ouïgour 13, but usage is not consistent throughout. In some documents, the terminals do not touch, as in U 320.

Pelliot ouïgour 13

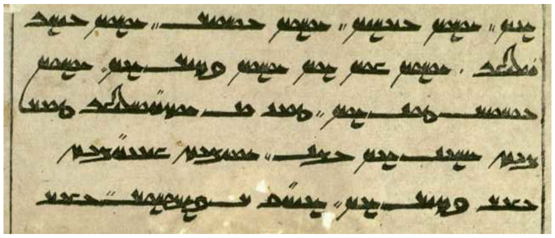


U 320

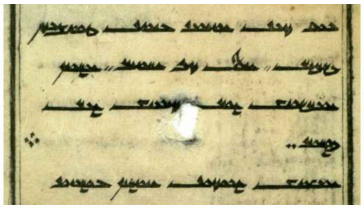


This technique is reproduced in some block-printed documents, such as U 388, that aim to represent the layout of the original 'square'-script document. But, it is not observed in the majority block prints, eg.

U 388

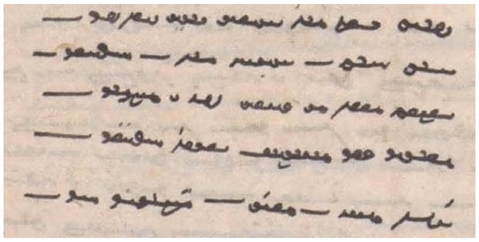


U 496

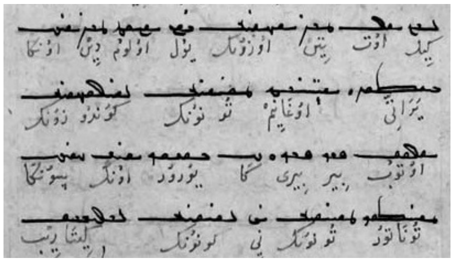


Terminal extension resulting in connections across words is not observed in later documents written in the ‘post-Mongolic’ style:

Kutadgu Bilig

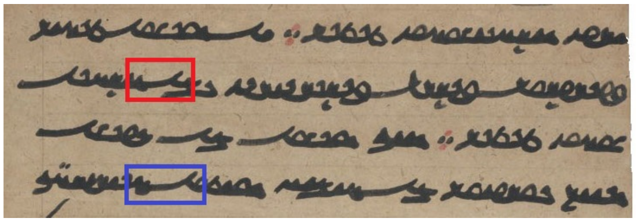


Atabetul Hakayik



This technique is a calligraphic practice and there is no semantic aspect to such inter-word connections. As observed in Pelliot ouïgour 13, above, the connections between words may be inconsistently produced. The precision of the connection has no bearing on the meaning of the text. The apparent linkage of words may have evolved as a function of cursive writing, as a way to maintain movement of the pen and ink, as a matter of convenience. Letters that have final shapes with elongated terminals naturally provide for swash strokes. But, as may be observed in the above specimens, final forms without long terminals are consistently not joined, as a matter of practice.

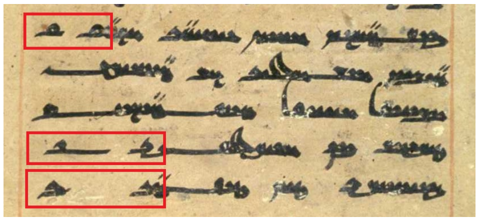
Furthermore, even when a terminal connects to a following word, the first letter of the latter is written using its initial form. This indicates that the behavior is stylistic, as observed in the below excerpt from Pelliot ouïgour 13, where an extended final *aleph* connects to the initial *aleph* of the following word, which is written using its distinctive initial form 𐽰 (highlighted red); and an extended final *nun* connects to the initial *gimel* of the following word, which is written using its distinctive initial form 𐽲 (highlighted blue):



In plain-text representation, the words joined by the terminal elongation should be separated using spaces. If a user wishes to represent the calligraphic appearance of the text as it appears on the page, the ZWNJ U+200C ZERO WIDTH NON-JOINER may be used instead of the space to effect a connection between words at the character level (see § 8.4). The actual rendering of the connection is to be handled typographically.

5.3 Initial stem extension

In addition to the elongation of terminals, another space-filling technique is leading initial baseline extension similar to *kashida*. If there is space between the last word on a line and the margin, the final letter of that word may be separated from the penultimate letter using an elongated baseline so that the space is filled by a 'bridge' between the letters. In some documents, the final letter of a word before the margin is reduplicated as a separate, isolated letter and prefixed with a baseline extension.



The Old Uyghur ـ 'stem extender' may be represented using ـ U+0640 ARABIC TATWEEL. The usage of the Arabic TATWEEL for Old Uyghur follows the Unicode convention of unifying stem extender characters in right-to-left scripts. The TATWEEL should be specified as a script extension for Old Uyghur as has been done for the Adlam, Hanifi Rohingya, Mandaic, Manichaean, Sogdian, and Syriac encodings (see § 9.3).

5.4 Terminal Orientation

As discussed above in the descriptions of letters, the terminals of *aleph*, *beth*, *nun*, etc. may have different orientations. There are possible explanations for such variation:

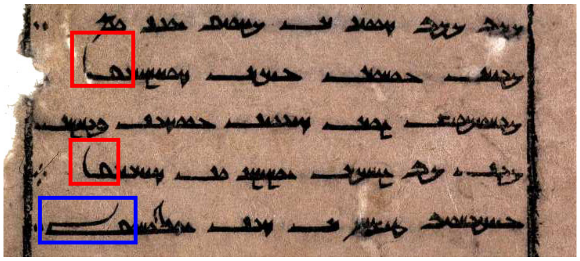
- *Spacing adjustment* When letters with vertical terminals occur at a margin with insufficient space to produce the regular stroke, the terminal is curved or hooked. In such cases, the direction of the tail has no semantic value.
- *Stylistic preference* In some documents written in a highly cursive style, a scribe may have a preference for the direction of terminals.
- *Intentional alternation* A scribe or block-printer may have explicitly chosen to use a variant terminal instead of the conventional form. This is apparent in the occurrence of both conventional and variant terminals at end of line, as well as in other position along a line. Intentional alternation is also evident in cases where both the conventional and variant forms are used simultaenously in a document in isolated contexts; this occurs frequently with *aleph*.

These alternate final forms are to be treated as glyphic variants. If a semantic difference between a variant and regular form is identified, then the variant form may be considered for encoding at that time.

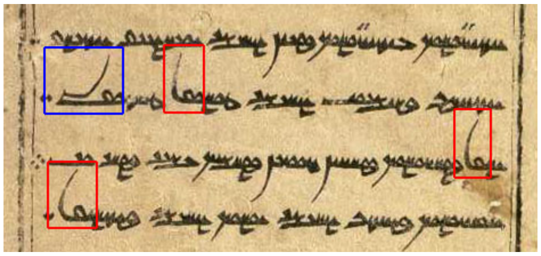
5.5 Space-filling terminal

A space-filling terminal is used in square and block-printed documents at the end of line. This terminal may represent the end of a section or a text. In the materials analyzed for this proposal, the terminal occurs exclusively with *pe*, eg. 𐽼 (blue) contrasted with the regular final form 𐽼 (red) in the excerpts below:

U 4750



U 4162

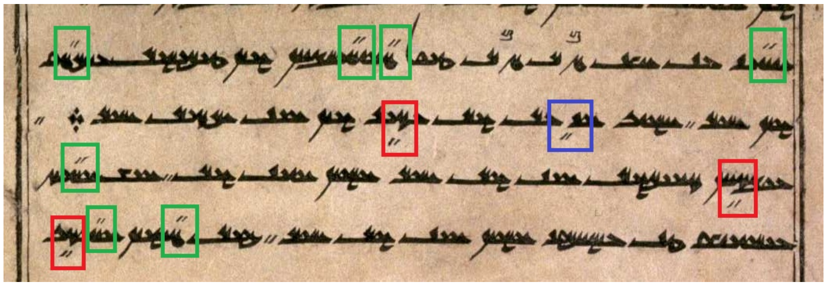


Von Gabain shows a sign in her chart of the script, annotated as “Zeilenfüller” (German “row-filler”), which resembles with a filled large dot instead of a loop (see fig. 5). It is not clear at present if is a stylistic variant of , or if the terminal is used with other letters. Additional research is required to determine the appropriate method for representing the terminal in encoded text. It has been added to the list in § 7 of characters not presently proposed.

5.6 Combining signs

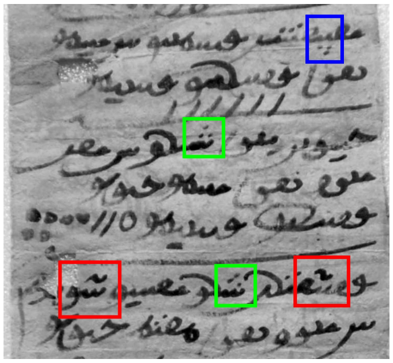
The following combining signs are commonly used in Old Uyghur documents used for disambiguation and representation of new sounds. Their usage with letters is described in § 8.2.

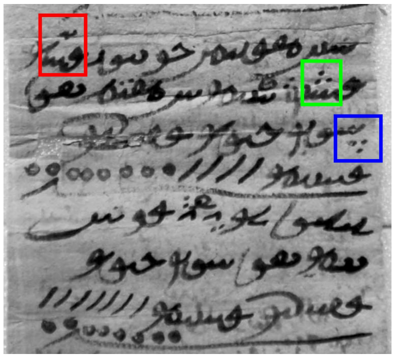
In Old Uyghur, dot diacritics are commonly used for differentiating between letters whose shapes are similar in particular styles of the script, and for indicating sounds for which distinctive letters do not exist in the script. These signs are commonly used with *nun*, *gimel*, *zayin*, *heth*, *samekh*, and *shin*.



The shape of these dot diacritics differ across the styles of the script. In the ‘square’ style and block prints, they are represented using elongated strokes, which reflect scribal aesthetics of the script. In the ‘cursive’ and later ‘miniscule’ style, these diacritics are written as true dots or squared dots. Despite the variations in their shapes, these signs are palaeographically dots, and therefore, it is appropriate to refer to them as such in the names for the proposed character.

In late Old Uyghur administrative documents, there are additional diacritics that are used for the transliteration of non-Turkic sounds, particularly those in Arabic words. Usage of the ◌̤ (blue), ◌̈ (green), and ◌̃̈ (red) for transcribing Arabic (from Israpil 2014: plate I).



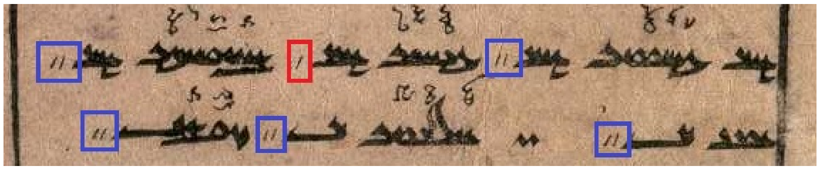


There are other signs used for similar purposes. Erdal (1984) describes the usage of other diacritics for diambiguation and transliteration of Arabic in administrative Old Uyghur documents of the 11th century from Yarkand. Clark (2010) also describes signs used in the *Kutadgu Bilig*, an 11th century Karakhanid work by Yusūf Khāṣṣ Ḥājib. A complete set of such diacritics will be proposed after additional research.

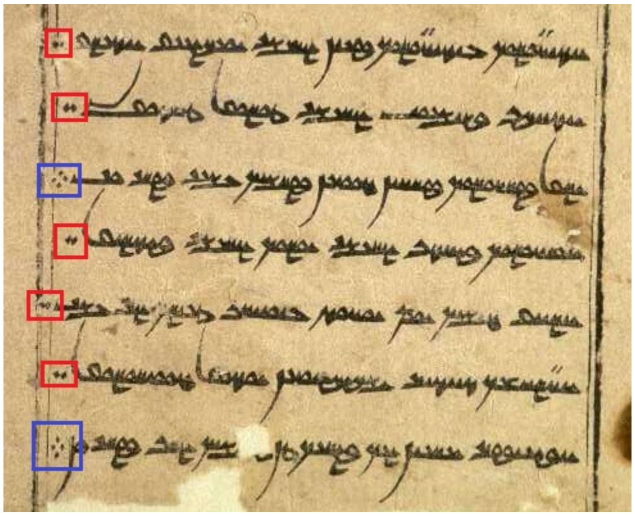
Both the traditional and later signs have the same semantic function as the *nuqṭa* diacritic, which is used in Brahmi-based scripts for representing sounds foreign to Indic languages, eg. ◌़ U+093C DEVANAGARI SIGN NUKTA. While it may be possible to encode combinations of base letter + combining sign as atomic letters, this approach should be avoided. There are other combining signs used in Old Uyghur manuscripts, which have not been fully investigated for the present proposal. It is quite likely that additional combining signs will need to be encoded. As a result, it will be necessary to encode new sets of atomic letters for each every base letter + combining sign combination when a new combining sign is added to the repertoire. The proposed model for combining signs follows that of the Sogdian encoding.

5.7 Punctuation signs

The signs 𐾆 and 𐾇 are common forms of punctuation (see Knüppel 2002). When used together, 𐾆 delimits shorter text segments, while 𐾇 indicates the end of longer segments, as in the excerpt from U 4123 below:

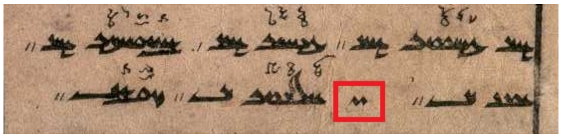


The signs 𐾈 and 𐾉 are used similarly, as in the excerpt from U 4162 below. While 𐾉 is similar to the generic ⁘ U+2058 FOUR DOT PUNCTUATION already encoded in Unicode, it should be encoded separately for Old Uyghur as it is part of a set of script-specific punctuation.

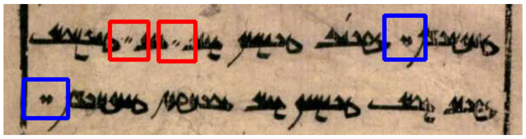


There is some variation in the form of 𐾈. As observed below, in U 4123 the ‘dots’ are not separated, but connected as to form a single sign resembling ∾. In U 7008 and U 343, the dots resemble ‘comma’-like shapes ❞ that touch at the bearings; note the regular shape of 𐾉 in U 343. At present, it is unknown if ∾ and ❞ are distinct forms of punctuations or glyphic variants of 𐾈, so they are not proposed for encoding.

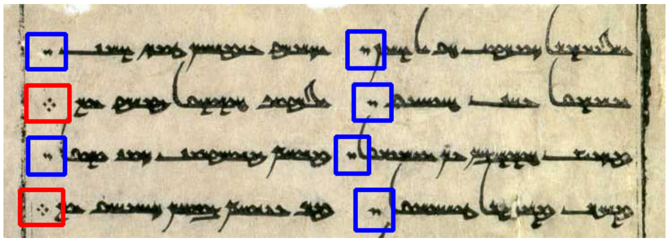
U 4123



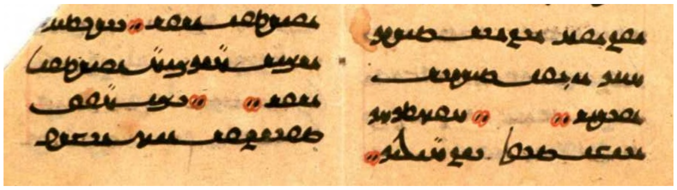
U 7008



U 343



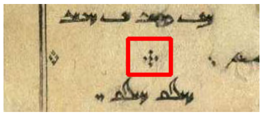
Punctuation resembling 𐫲 appears in Old Uyghur manuscripts. This sign should be unified with 𐫲 U+10AF2 MANICHAEAN PUNCTUATION DOUBLE DOT WITHIN DOT and specified as a script extension (see § 9.3).



The ꝁ is shown as a sign of punctuation in the list of characters used in the inscriptions on the walls of the Cloud Platform at Juyong Guan (see fig. 6). Additional research is required to understand its usage and suitability for encoding.

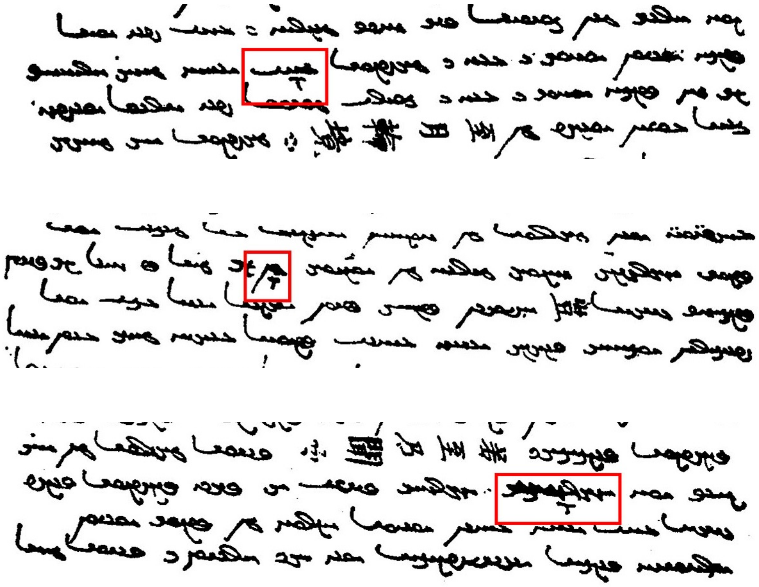
The ᐟ•• and ❞ occur in some documents (see § 5.7), but, it is unknown if these signs are distinct forms of punctuations or glyphic variants of 𐾈. They may be unified with 𐾈 at present. If additional research indicates that they are distinctive signs of punctuation, they may be proposed for encoding in the future.

The 𐾉 is used as a sign of punctuation and decoration in U 4124. If additional attestations are identified, it may be encoded as part of the block.



5.8 Editorial sign

When written beneath a word or letter, the $\underset{\top}{\circ}$ deletion sign indicates that the respective text is an error and is to be omitted. In authentic representations of manuscripts, it is to be placed after the letter that carries the mark. The correct word is generally written after the mispelled word. Usage of the $\underset{\top}{\circ}$ deletion mark for indicating error correction in Or. 8212/75, an Old Uyghur manuscript containing passages of the of the Buddhist text *Abhidharma-nyāyānusāra-śāstra* (from Shōgaito 1988: 207). Note the intralinear text in Han characters.



The deletion sign is suitable for encoding, but additional research to identify other editorial signs used in Old Uyghur should be conducted in order to determine usage frequencies. If additional editorial characters are identified, the deletion sign and the others may be proposed together for encoding.

6 Encoding Model

6.1 Script block

The proposed 'Old Uyghur' script block is allocated to four columns in the Supplementary Multilingual Plane (SMP) beginning at the code point U+10F70. The repertoire proposed at present contains 26 characters. The remaining code points are required for numerous characters that have been identified, but which will be proposed for encoding at a later time (see § 7).

6.2 Scope of the encoding

The proposed encoding enables representation of typical Old Uyghur documents. The character inventory is based upon the alphabet attested in the 9th century manuscript U 40 and the 11th century treatise by Kāšġarī. The inventory is confirmed by palaeographic analysis of isolated, initial, medial, and final forms of letters attested in Old Uyghur manuscripts in the square, semi-square, semi-cursive, and cursive styles, as well as manuscript facsimilies reproduced in metal type in Müller (1908).

6.3 Representative glyphs

Representative glyphs are based upon the 'square' style of the Old Uyghur script. This style is the traditional hand that conveys the most distinctions between letters and was used for producing formal documents. It is also the basis for block-printing types. Isolated forms of letters are based upon U 40 and Kāšġarī, which convey a tradition of using these forms in enumerations of the alphabet. This differs from the practice in Unicode of using final forms as the isolated letters for cursive joining scripts. Contextual forms of the letters are based upon normalizations of shapes used in the 'square' style and in block prints, and validated using the printed forms in Müller (1908).

6.4 Unification of styles

The proposed encoding unifies all styles of the Old Uyghur script. While the representative glyphs are based on the 'square' style, fonts may be created for rendering the script in other styles.

6.5 Character names

The names of Old Uyghur letters are based upon scholarly names for the original Sogdian letters, which in turn reflect the ancestral Aramaic names.

- Throughout this proposal, italics are used for scholarly names for graphemes, while small capitals indicate Unicode character names, eg. 𐽰 is referred to as the grapheme *aleph* and the proposed Unicode character is formally referred to as OLD UYGHUR LETTER ALEPH. For brevity, in this document, when referring to a proposed Unicode character, the descriptor 'OLD UYGHUR' and the character class, eg. 'LETTER', may be dropped, eg. OLD UYGHUR LETTER ALEPH is truncated to ALEPH. Characters of other scripts are designated by their full Unicode names. Latin transliteration of Old Uyghur follows the current scholarly convention.

- The descriptors 'above' and 'below' in the character names refer to the orientation of features with respect to the horitonzal baseline of the script. In vertical contexts, 'above' should be interpreted as 'left', and 'below' as 'right'.

6.6 Proposed encoded character repertoire

The proposed encoding repertoire contains 26 characters (the code chart and names list follows p. 11):

- 18 letters, which represent all palaeographically distinct letters
- 4 combining signs for representing traditional diacritics
- 4 punctuation signs

Letters

Character name	Glyph	Joining	Latin
OLD UYGHUR LETTER ALEPH	𐽰	dual	ʾ
OLD UYGHUR LETTER BETH	𐽱	dual	β
OLD UYGHUR LETTER GIMEL-HETH	𐽲	dual	γ, x, q
OLD UYGHUR LETTER WAW	𐽳	dual	w
OLD UYGHUR LETTER ZAYIN	𐽴	right	z, ž
OLD UYGHUR LETTER FINAL HETH	𐽵	right	-x, -q
OLD UYGHUR LETTER YODH	𐽶	dual	y
OLD UYGHUR LETTER KAPH	𐽷	dual	k
OLD UYGHUR LETTER LAMEDH	𐽸	dual	δ
OLD UYGHUR LETTER MEM	𐽹	dual	m
OLD UYGHUR LETTER NUN	𐽺	dual	n
OLD UYGHUR LETTER SAMEKH	𐽻	dual	s
OLD UYGHUR LETTER PE	𐽼	dual	p
OLD UYGHUR LETTER SADHE	𐽽	dual	c
OLD UYGHUR LETTER RESH	𐽾	dual	r
OLD UYGHUR LETTER SHIN	𐽿	dual	š

OLD UYGHUR LETTER TAW	𐾀	dual	t
OLD UYGHUR LETTER LESH	𐾁	dual	l

Combining signs

Character name	Glyph
OLD UYGHUR COMBINING DOT ABOVE	◌𐾂
OLD UYGHUR COMBINING DOT BELOW	◌𐾃
OLD UYGHUR COMBINING TWO DOTS ABOVE	◌𐾄
OLD UYGHUR COMBINING TWO DOTS BELOW	◌𐾅

Punctuation signs

Character name	Glyph
OLD UYGHUR PUNCTUATION BAR	𐾆
OLD UYGHUR PUNCTUATION TWO BARS	𐾇
OLD UYGHUR PUNCTUATION TWO DOTS	𐾈
OLD UYGHUR PUNCTUATION FOUR DOTS	𐾉

6.7 Collation

The sort order for Old Uyghur letters follows the encoded order:

𐽰 ALEPH < 𐽱 BETH < 𐽲 GIMEL-HETH < 𐽳 WAW < 𐽴 ZAYIN < 𐽵 FINAL HETH < 𐽶 YODH < 𐽷 KAPH < 𐽸 LAMEDH < 𐽹 MEM < 𐽺 NUN < 𐽻 SAMEKH < 𐽼 PE < 𐽽 SADHE < 𐽾 RESH < 𐽿 SHIN < 𐾀 TAW < 𐾁 LESH

6.8 Contextual forms of letters

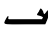
Contextual forms of Old Uyghur letters are shown below:

	joining	X_n	X_f	X_m	X_i
ALEPH	dual	𐽰	‍𐽰	‍𐽰‍	𐽰‍
BETH	dual	𐽱	‍𐽱	‍𐽱‍	𐽱‍
GIMEL-HETH	dual	𐽲	‍𐽲	‍𐽲‍	𐽲‍
FINAL HETH	right	𐽵	‍𐽵	—	—
WAW	dual	𐽳	‍𐽳	‍𐽳‍	𐽳‍
ZAYIN	right	𐽴	‍𐽴	—	—
YODH	dual	𐽶	‍𐽶	‍𐽶‍	𐽶‍
KAPH	dual	𐽷	‍𐽷	‍𐽷‍	𐽷‍
LAMEDH	dual	𐽸	‍𐽸	‍𐽸‍	𐽸‍
MEM	dual	𐽹	‍𐽹	‍𐽹‍	𐽹‍
NUN	dual	𐽺	‍𐽺	‍𐽺‍	𐽺‍
SAMEKH	dual	𐽻	‍𐽻	‍𐽻‍	𐽻‍
PE	dual	𐽼	‍𐽼	‍𐽼‍	𐽼‍
SADHE	dual	𐽽	‍𐽽	‍𐽽‍	𐽽‍
RESH	dual	𐽾	‍𐽾	‍𐽾‍	𐽾‍
SHIN	dual	𐽿	‍𐽿	‍𐽿‍	𐽿‍
TAW	dual	𐾀	‍𐾀	‍𐾀‍	𐾀‍
LESH	dual	𐾁	‍𐾁	‍𐾁‍	𐾁‍

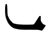
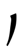
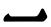
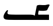
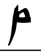
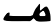
7 Characters Not Proposed for Encoding

The characters described in this section are not included in the repertoire proposed at present. They will be proposed for inclusion at a later time after additional research.

Alternate forms (see § 5.1.1)

Description	Alternate	Regular
‘toothed’ *aleph*		
‘toothed’ *aleph* with upward terminal		

Letters with terminal variants (see § 5.4)

Description	Variant	Regular
aleph with upward terminal		
aleph with downward terminal		
beth with upward terminal		
kaph with upward terminal		
nun with downward terminal		
sadhe with downward terminal		
taw with downward terminal		

Space-filling terminal (see § 5.5)

Description	Glyph	Joining
space-filling terminal		right

Combining signs (see § 5.6)

Description	Glyph
combining three dots above	◌ۛ
combining three dots below	
combining *hamza* above	◌ٔ
combining ring above	◌̊
combining ring below	◌̥

Punctuation (see § 5.7)

Description	Glyph
five-dot punctuation	⁙
Cloud Platform section mark	
connected dots	
‘comma’-like dots	

Editoral sign (see § 5.8)

Description	Glyph
deletion sign	

8 Encoded representations

8.1 Vowels

The representation of vowels follows the ‘*matres lectionis*’ pattern for Semitic scripts, in which 𐽰 *aleph*, 𐽳 *waw*, and 𐽶 *yodh* are used for indicating vowels. These letters are combined in digraphs and trigraphs in order to express the vowel repertoire of Turkic languages, as shown below:

	Word-initial		Word-medial	
ä	𐽰	𐽰 ALEPH	𐽰	𐽰 ALEPH
a, *e*	𐽰𐽰	𐽰 ALEPH, 𐽰 ALEPH	𐽰	𐽰 ALEPH
i, *ï*	𐽰𐽶	𐽰 ALEPH, 𐽶 YODH	𐽶	𐽶 YODH
ī, *ï̄*	𐽰𐽶𐽶	𐽰 ALEPH, 𐽶 YODH, 𐽶 YODH	𐽶𐽶	𐽶 YODH, 𐽶 YODH
o, *u*	𐽰𐽳	𐽰 ALEPH, 𐽳 WAW	𐽳	𐽳 WAW
ö, *ü*	𐽰𐽳𐽶	𐽰 ALEPH, 𐽳 WAW, 𐽶 YODH	𐽳	𐽳 WAW
ö, *ü*	𐽳𐽶	𐽳 WAW, 𐽶 YODH	𐽳𐽶	𐽳 WAW, 𐽶 YODH
ō, *ȫ*, *ū*, *ǖ*	𐽰𐽳𐽳	𐽰 ALEPH, 𐽳 WAW, 𐽳 WAW	𐽳𐽳	𐽳 WAW, 𐽳 WAW

8.2 Disambiguation and extension of letters

The combining signs enumerated in § 5.6 are written with letters to diambiguate consonants or to represent consonants for which distinctive letters do not exist. The following forms are attested. Combining signs are placed after a letter in encoded text:

		X_n	X_f	X_m	X_i	
dotted *gimel, heth*	*γ*	𐽲𐾂	𐽲𐾂	𐽲𐾂	𐽲𐾂	𐽲 GIMEL-HETH, ◌𐾂 COMBINING DOT ABOVE
two-dotted *gimel, heth*	*γ*	𐽲𐾄	𐽲𐾄	𐽲𐾄	𐽲𐾄	𐽲 GIMEL-HETH, ◌𐾄 COMBINING TWO DOTS ABOVE
dotted *zayin*	*ž*	𐽴𐾃	𐽴𐾃	—	—	𐽴 ZAYIN, ◌𐾃 COMBINING DOT RIGHT
two-dotted *zayin*	*ž*	𐽴𐾅	𐽴𐾅	—	—	𐽴 ZAYIN, ◌𐾅 COMBINING TWO DOTS RIGHT

dotted *heth*	*q*	𐽵𐾂	𐽵𐾂	—	—	𐽵 FINAL HETH, ◌𐾂 COMBINING DOT ABOVE
two-dotted *heth*	*q*	𐽵𐾄	𐽵𐾄	—	—	𐽵 FINAL HETH, ◌𐾄 COMBINING TWO DOTS ABOVE
dotted *nun*	*n*	𐽺𐾂	𐽺𐾂	𐽺𐾂	𐽺𐾂	𐽺 NUN, ◌𐾂 COMBINING DOT ABOVE
two-dotted *shin*	*š*	𐽿̤	𐽿̤	𐽿̤	𐽿̤	𐽿 SHIN, ◌̤ COMBINING TWO DOTS RIGHT

8.3 Stem extension

Stem extension is to be represented in encoded text using ـ U+0640 ARABIC TATWEEL:

tynlγ lr-r ‘tinlag-lar-r’	𐾀𐽶𐽺𐾁𐽲 𐾁𐽾 ـ𐽾	𐾀 TAW, 𐽶 YODH, 𐽺 NUN, 𐾁 LESH, 𐽲 GIMEL, SP SPACE, 𐾁 LESH, 𐽾 RESH, SP SPACE, ـ U+0640 ARABIC TATWEEL, 𐽾 RESH

8.4 Terminal connections

As described in § 5.2, a letter with an elongated terminal may be written so as to touch the initial letter of the following word. This calligraphic or stylistic technique may be reproduced in encoded text using ZWNJ ZWNJ instead of a space:

yma ʾʾγyn yma algin	𐽶𐽹𐽰 𐽰𐽰𐾁𐽲𐽶𐽺	𐽶 YODH, 𐽹 MEM, 𐽰 ALEPH, SP SPACE, 𐽰 ALEPH, 𐽰 ALEPH, 𐾁 LESH, 𐽲 GIMEL, 𐽶 YODH, 𐽺 NUN
	𐽶𐽹𐽰‌𐽰𐽰𐾁𐽲𐽶𐽺	𐽶 YODH, 𐽹 MEM, 𐽰 ALEPH, ZWNJ ZWNJ, 𐽰 ALEPH, 𐽰 ALEPH, 𐾁 LESH, 𐽲 GIMEL, 𐽶 YODH, 𐽺 NUN

8.5 Handling Ambiguity

The encoding for Old Uyghur does not aim to, nor could it be expected to, provide a means for representing all ambiguous readings that result from indistinct, cursive, or rapid writing. A Unicode encoding cannot attempt to account for idiosyncratic scribal practices that result in ambiguous readings. Indecipherability of a piece of text is not so much a problem of what is written — the underlying text was likely written to communicate specific meaning and may have been comprehensible to a reader familiar with the styles and nuances of writing during that time — but it is a problem for a modern reader who is unfamiliar with the script or underlying language.

It is unreasonable to expect that a Unicode encoding would fully enable a user attempting to transcribe a piece of text without knowing how to distinguish one letter from the other, without knowing the underlying language or rules of the script, whether it is Old Uyghur, Latin, Cyrillic, Arabic, or any other script with quite dynamic cursive traditions. The natural ambiguity of a cursive text that could be read in multiple ways by a person unfamiliar with the language, might probably be deciphered quickly by someone familiar with the language through morphological and syntactic or other linguistic contexts.

In any case, when representing Old Uyghur text in which the the identity of a character cannot be established, the user should follow a strategy that entails:

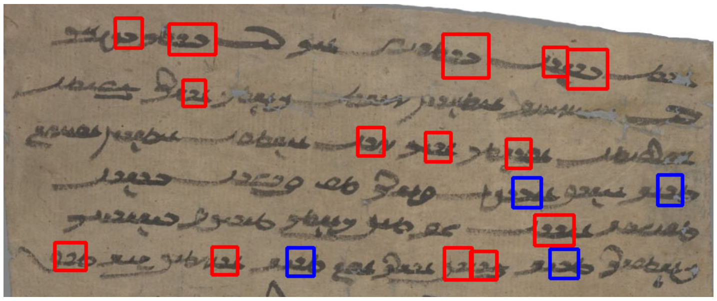
1. Analyze pairs of letters, such as *aleph* and *nun*, or *beth* and *yodh*, or *samekh* or *shin* that are consistently confounded in order to identify the degree to which the letters of the pair are similar.

2. If the pair is indistinct, the user should select the encoded character whose glyph most closely resembles the shape of the letters in the text. That character should be used consistently throughout the text for representing all instances of the letters. For example, in texts where /s/ and /š/ are written using a single letter, select either the character SAMEKH or SHIN, based upon which letter most closely resembles the form in the text, and use that letter for all occurrences in the text.

3. Document the character used for ambigious readings, along with the rationale for selecting the particular character.

4. If there is a need to display the text as it appears in the original source, a font should be developed using glyphs that match the original.

For example, as shown in § 5.1.2, in some less carefully written documents, *beth* and *yodh* may be written using an ambiguous shape, eg. 𐽶, that approximates the general outline of the two letters. This is evident in Pelliot ouïgour 2, where there is a lack of consistency in distinguishing *beth* (blue) from *yodh* (red):



The shape of *yodh* in this document varies considerably. Sequences of the letter do not have consistent shapes. For example, the initial letters of the second and third words of line 1 are *yodh*. However, the shape of initial *yodh* in the second word is more open than the angular form of initial *yodh* in the third word. On account of this, a user unfamiliar with Old Uyghur might interpret the initial letter of the second word as 𐽱 *beth* instead of 𐽶 *yodh*. Similarly, the medial *beth* in the first word of line 4 could be construed as a *yodh* based upon the variance of *yodh* in the preceding lines. If the non-specialist user is unable to distinguish *beth* and *yodh* from contextual clues, they should choose to represent all instances of these letters using either BETH or YODH. This approach may not preserve the underlying text with complete accuracy, but given that these two letters have the same properties, it would provide a means for displaying the text to an acceptable degree.

9 Character Properties

9.1 Core data: `UnicodeData.txt`

```
10F70;OLD UYGHUR LETTER ALEPH;Lo;0;R;;;;;N;;;;;
10F71;OLD UYGHUR LETTER BETH;Lo;0;R;;;;;N;;;;;
10F72;OLD UYGHUR LETTER GIMEL-HETH;Lo;0;R;;;;;N;;;;;
10F73;OLD UYGHUR LETTER WAW;Lo;0;R;;;;;N;;;;;
10F74;OLD UYGHUR LETTER ZAYIN;Lo;0;R;;;;;N;;;;;
10F75;OLD UYGHUR LETTER FINAL HETH;Lo;0;R;;;;;N;;;;;
10F76;OLD UYGHUR LETTER YODH;Lo;0;R;;;;;N;;;;;
10F77;OLD UYGHUR LETTER KAPH;Lo;0;R;;;;;N;;;;;
10F78;OLD UYGHUR LETTER LAMEDH;Lo;0;R;;;;;N;;;;;
10F79;OLD UYGHUR LETTER MEM;Lo;0;R;;;;;N;;;;;
10F7A;OLD UYGHUR LETTER NUN;Lo;0;R;;;;;N;;;;;
10F7B;OLD UYGHUR LETTER SAMEKH;Lo;0;R;;;;;N;;;;;
10F7C;OLD UYGHUR LETTER PE;Lo;0;R;;;;;N;;;;;
10F7D;OLD UYGHUR LETTER SADHE;Lo;0;R;;;;;N;;;;;
10F7E;OLD UYGHUR LETTER RESH;Lo;0;R;;;;;N;;;;;
10F7F;OLD UYGHUR LETTER SHIN;Lo;0;R;;;;;N;;;;;
10F80;OLD UYGHUR LETTER TAW;Lo;0;R;;;;;N;;;;;
10F81;OLD UYGHUR LETTER LESH;Lo;0;R;;;;;N;;;;;
10F82;OLD UYGHUR COMBINING DOT ABOVE;Mn;230;NSM;;;;;N;;;;;
10F83;OLD UYGHUR COMBINING DOT BELOW;Mn;220;NSM;;;;;N;;;;;
10F84;OLD UYGHUR COMBINING TWO DOTS ABOVE;Mn;230;NSM;;;;;N;;;;;
10F85;OLD UYGHUR COMBINING TWO DOTS BELOW;Mn;220;NSM;;;;;N;;;;;
10F86;OLD UYGHUR PUNCTUATION BAR;Po;0;R;;;;;N;;;;;
10F87;OLD UYGHUR PUNCTUATION TWO BARS;Po;0;R;;;;;N;;;;;
10F88;OLD UYGHUR PUNCTUATION TWO DOTS;Po;0;R;;;;;N;;;;;
10F89;OLD UYGHUR PUNCTUATION FOUR DOTS;Po;0;R;;;;;N;;;;;
```

9.2 Linebreak data: `LineBreak.txt`

```
10F70..10F81;AL # Lo [18] OLD UYGHUR LETTER ALEPH..OLD UYGHUR LETTER LESH
10F82..10F85;CM # Mn  [4] OLD UYGHUR COMBINING DOT ABOVE..
                          OLD UYGHUR COMBINING TWO DOTS BELOW
10F86..10F89;AL # Po  [4] OLD UYGHUR PUNCTUATION BAR..OLD UYGHUR PUNCTUATION FOUR DOTS
```

9.3 Script extensions: `ScriptExtensions.txt`

```
# Script_Extensions=Adlm Arab Mand Mani Phlp Rohg Sogd Syrc

0640          ; Adlm Arab Mand Mani Phlp Rohg Sogd Syrc # Lm       ARABIC TATWEEL

# Total code points: 1

# Script_Extensions=Mani

10AF2        ; Mani # Po       MANICHAEAN PUNCTUATION DOUBLE DOT WITHIN DOT

# Total code points: 1
```

9.4 Shaping properties: `ArabicShaping.txt`

```
10F70; OLD UYGHUR ALEPH; D; No_Joining_Group
10F71; OLD UYGHUR BETH; D; No_Joining_Group
10F72; OLD UYGHUR GIMEL-HETH; D; No_Joining_Group
10F73; OLD UYGHUR WAW; D; No_Joining_Group
10F74; OLD UYGHUR ZAYIN; D; No_Joining_Group
10F75; OLD UYGHUR FINAL HETH; R; No_Joining_Group
10F76; OLD UYGHUR YODH; D; No_Joining_Group
10F77; OLD UYGHUR KAPH; D; No_Joining_Group
10F78; OLD UYGHUR LAMEDH; D; No_Joining_Group
10F79; OLD UYGHUR MEM; D; No_Joining_Group
10F7A; OLD UYGHUR NUN; D; No_Joining_Group
10F7B; OLD UYGHUR SAMEKH; D; No_Joining_Group
10F7C; OLD UYGHUR PE; D; No_Joining_Group
10F7D; OLD UYGHUR SADHE; D; No_Joining_Group
10F7E; OLD UYGHUR RESH; D; No_Joining_Group
10F7F; OLD UYGHUR SHIN; D; No_Joining_Group
10F80; OLD UYGHUR TAW; D; No_Joining_Group
10F81; OLD UYGHUR LESH; D; No_Joining_Group
```

10 References

Anderson, Deborah, et. al. 2018. “Recommendations to UTC #155 April-May 2018 on Script Proposals” (L2/18-168). `https://www.unicode.org/L2/L2018/18168-script-rec.pdf`

———. 2019. “Recommendations to UTC #158 January 2019 on Script Proposals” (L2/19-047). `https://www.unicode.org/L2/L2019/19047-script-adhoc-recs.pdf`

———. 2020. “Recommendations to UTC #162 January 2020 on Script Proposals” (L2/20-046). `https://www.unicode.org/L2/L2020/20046-script-adhoc-rept.pdf`

China. 2018. “GB/T 36331-2018 ”Information technology – Uigur-Mongolian characters, presentation characters and use rules of controlling characters”. `http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=DFE87CC79EA67F8BF8B37C9C41CF9348`

Clark, Larry. 2010. “The Turkic script and the *Kutadgu Bilig*”. *Turcology in Mainz*, Turcologica, Band 82, Hendrik E Boeschoten and Julian Rentzsch (ed.), pp. 89–106. Wiesbaden: Harrassowitz Verlag.

Clauson, Gerard. 1962. *Studies in Turkic and Mongolic Linguistics*. London: Royal Asiatic Society of Great Britain and Ireland.

Coulmas, Florian. 1996. *The Blackwell Encyclopedia of Writing Systems*. Oxford: Blackwell Publishers.

Erdal, Marcel. 1984. “The Turkish Yarkand Documents”. *Bulletin of the School of Oriental and African Studies, University of London*, vol. 47, no. 2, pp. 260–301.

Hamilton James R. 1986. *Manuscrits ouïgours du IXè-Xè siècle de Touen-Houang*: textes établis, traduits, et commentés. Tome 1. Paris.

Israpil, Dilara; Yüsüp, Israpil. 2014. “Two Old Uighur Account Documents from Toqquzsaray Ruins in Maralbeši”. *Studies on the Inner Asian Languages*, vol. 29, pp. 137–156.

Kara, György. 1996. "Aramaic Scripts for Altaic Languages". *The World's Writing Systems*, Peter T. Daniels and William Bright (ed.), pp. 536–558. New York and Oxford: Oxford University Press.

Klaproth, Heinrich Julius. 1812. *Abhandlung über die Sprache und Schrift der Uiguren*. Berlin.

Knüppel, Michael. 2002. "Zu den „Daṇḍas" und „Doppel-Daṇḍas" in der uigurischen Schrift". *Acta Orientalia Academiae Scientiarum Hungaricae*, vol. 55, no. 4, pp. 339–343.

Kontovas, Nicholas. 2020. "Endorsement of the Old Uyghur encoding proposal L2/20-191". L2/20-199. `https://www.unicode.org/L2/L2020/20199-old-uyghur-support.pdf`

Le Coq, Albert von. 1919. "Kurze Einführung in die uigurische Schriftkunde". *Mitteilungen des Seminars für Orientalische Sprachen an der Friedrich-Wilhelms-Universität zu Berlin*, vol. 22, pt. 2, pp. 93–109.

Mair, Victor. 2009. "A Little Primer of Xinjiang Proper Nouns". `https://languagelog.ldc.upenn.edu/nll/?p=1576`

Matsui, Dai. 2017. "An Old Uigur Account Book for Manichaean and Buddhist Monasteries from Temple α in Qočo". *Zur lichten Heimat: Studien zu Manichäismus, Iranistik und Zentralasienkunde im Gedenken an Werner Sundermann*, Herausgegeben von einem Team „Turfanforschung", pp. 409–420. Wiesbaden: Harrassowitz Verlag.

———. 2018. "Comments on the preliminary proposal to encode Old Uyghur in Unicode (L2/18-126)". L2/18-335. `https://www.unicode.org/L2/L2018/18335-old-uyghur-cmt.pdf`

Moriyasu, Takao. 1989. "Notes on Uighur Documents (I)." *Studies on the Inner Asian Languages*, vol. 4, pp. 51–76.

———. 2004. "From Silk, Cotton and Copper Coin to Silver. Transition of the Currency Used by the Uighurs during the Period from the 8th to the 14th Centuries". *Turfan Revisited: The First Century of Research into the Arts and Cultures of the Silk Road*, Desmond Durkin-Meisterernst, et al. (eds.), pp. 228–239. Berlin: Dietrich Reimer Verlag.

Müller, Friedrich Wilhelm Karl. 1908. *Uigurica*. Abhandlungen der Königlich Preußischen Akademie der Wissenschaften, Philosophisch-historische Klasse 2. Berlin: Verlag der Königlichen Akademie der Wissenschaften.

———. 1910. *Uigurica*, vol. II. Berlin: Verlag der Königlichen Akademie der Wissenschaften.

Nadeliaev, V. M.; Nasilov, D. M.; Scherbak, A. M.; Tenishev, E. R. 1969. *Drevnetiurkskii slovar*. Akademiia nauk SSSR. Institut iazykoznaniia. Leningrad: Izd-vo "Nauka" Leningradskoe otd-nie.

Ölmez, Mehmet. 2016. "Compared transcription system for Old Uyghur Alphabet". Lecture at École des Hautes Études en Sciences Sociales (ÉHÉSS), Paris, May 2016. `http://www.academia.edu/24939281/Lectures_at_EHESS_1_Compared_transcription_system_for_Old_Uighur_Alphabet`

Osman, Omarjan. 2012. "Proposal for encoding the Uygur script in the SMP". L2/12-066. `https://www.unicode.org/L2/L2012/12066-uygur.pdf`

———. 2013. “Proposal to Encode the Uyghur Script in ISO/IEC 10646”. L2/13-071. `http://www.unicode.org/L2/L2013/13071-uyghur.pdf`

Pandey, Anshuman. 2016. “Revised proposal to encode the Sogdian script in Unicode”. L2/16-371R2. `http://www.unicode.org/L2/L2016/16371r2-sogdian.pdf`

Radloff, W [Radlov, Vasiliĭ Vasil′evich]. 1908. “Die vorislamitisehen Schriftarten der Türken und ihr Verhältniss zu der Sprache derselben”. *Bulletin de l’Académie Impériale des Sciences de St.-Pétersbourg*, pp. 835–856.

———. 1910. *Ṭišastvustik: Ein in Türkischer Sprache bearbeitetes Buddhistisches Sūtra*. I. Transcription and Übersetzung; II. Bemerkungen zu den Brāhmīglossen des Ṭišastvustik-Manuscripts (Mus. As. Kr. VII) von Baron A. von Staël-Holstein. Bibliotheca Buddhica, XII. St.-Pétersbourg.

———. 1911. *Kuan-ši-im Pusar: Eine türkische Übersetzung des XXV. Kapitels der chinesischen Ausgabe des Saddharmapuṇḍarīka*. Bibliotheca Buddhica XIV. St.-Pétersbourg: Commissionnaires de l’Académie Impériale des Sciences.

Radlov, Vasiliĭ Vasil′evich; Malov, S. Efimovich. 1913. *Suvarṇaprabhāsa*. (Sutra zolotogo bleska); tekst ujgurskogo redakcii. Bibliotheca Buddhica, XVII. Sanktpeterburg: Imper. Akad. Nauk.

Ščerbak, A. M. 1982. “De l’alphabet ouigour”. *Acta Orientalia Academiæ Scientiarum Hungaricæ*, vol. 36, no. 1–3, pp. 469–474.

Shōgaito, Masahiro. 1988. “Passages from Abhidharma-nyāyānusāra-śāstra. Quoted in the Uighur Text Or. 8212-75B, British Library”. *Studies on the Inner Asian Languages*, vol. 3, pp. 159–207.

Sims-Williams, Nicholas. 1981. “The Sogdian sound-system and the origins of the Uyghur script”. *Journal Asiatique*, pp. 347–360.

Terminology Normalization Committee for Ethnic Languages of the Xinjiang Uyghur Autonomous Region. 2006. “Recommendation for English transcription of the word ئۇيغۇر/ 《维吾尔》 ” (11 October 2006).

von Gabain, Annemarie. 1950. *Alttürkische Grammatik*. Mit Bibliographie, Lesestücken und Wörterverzeichnis, auch Neutürkisch. Mit vier Schrifttafeln under sieben Schriftproben. Porta linguarum orientalium. no. 23. 1. verbesserte Auflage. Leipzig: Otto Harrassowitz.

West, Andrew. 2006. “Phags-pa Script : Old Uyghur Script”. `http://www.babelstone.co.uk/Phags-pa/Uighur.html`

———. 2011a. “Khitan Miscellanea 1: Oh, How the Gods Mock Us!”. `http://babelstone.blogspot.com/2011/10/khitan-miscellanea-1.html`

———. 2011b. “Phags-pa Uyghur Seals”. `http://babelstone.blogspot.com/2011/11/phags-pa-uyghur-seals.html`

Wilkens, Jens. 2016. “Buddhism in the West Uyghur Kingdom and Beyond”. *Transfer of Buddhism Across Central Asian Networks (7th to 13th Centuries)*, Carmen Meinert (ed.), pp. 191–249. Brill: Leiden and Boston.

Yakup, Abdurishid. 2011. "An Old Uyghur fragment of the Lotus Sūtra from the Krotkov collection in St. Petersburg." *Acta Orientalia Academiæ Scientiarum Hungaricæ*, vol. 64, no. 4 (December), pp. 411–426.

Zieme, Peter. 1975. "Zur Buddhistischen Stabreimdichtung der alten Uiguren". *Acta Orientalia Academiæ Scientiarum Hungaricæ*, vol. 29, no. 2, pp. 187–211.

———. 1991. *Die Stabreimtexte der Uiguren von Turfan und Dunhuang: Studien zur alttürkischen Dichtung*. Bibliotheca orientalis Hungarica, v. XXXIII. Budapest: Akadémiai Kiadó.

11 Acknowledgments

I express my gratitude to:

- Dr. Dai Matsui (Graduate School of Letters, Osaka University) for sharing his wealth of knowledge about the Old Uyghur script, particularly its historical evolution, orthographic intracacies, and stylistic variations. I truly appreciate him for providing samples of manuscripts and for pointing me to specific examples of letterforms and orthographic issues.

- Dr. Mehmet Ölmez (Department of Modern Turkic Languages and Literatures, Istanbul University) for sharing excerpts of his book with me, as well as images of the *Dīwān luġāt al-turk* of Kāšġarī, and several other publications.

- Dr. Yukio Kasai (Centrum für Religionswissenschaftliche Studien, Ruhr-Universität Bochum) for providing details on the classification of Old Uyghur styles.

- Nicholas Kontovas (Leiden University) for providing documentation of the Old Uyghur script used in post-Mongolic documents. He has also patiently answered my numerous questions about the styles and usage of the script.

- Liang Hai for providing extensive comments on all previous versions of this proposal. The present document owes its shape to his meticulous review and recommendations.

- Roozbeh Pournader for providing invaluable comments and suggestions regarding the representation of right-to-left scripts. Not only this proposal, but every proposal I've submitted, and by extension every script I've encoded in Unicode, would not have been possible without his expertise and insight.

- Andrew West for providing reproductions and transcriptions of the Old Uyghur inscriptions on the walls of the Cloud Platform at Juyong Guan.

- The Script Ad Hoc group of the Unicode Technical Committee for reviewing all versions of this proposal and providing extensive feedback.

Digitized images of Old Uyghur documents from the Berlin Turfan Collection and the International Dunhuang Project have been used in accordance with academic fair-use conventions. I express my gratitude to Berlin-Brandenburgischen Akademie der Wissenchaften (Staatsbibliotek zu Berlin, Preussischer Kulturbesitz Orientabteilung) and to the British Library for making these images available. Documents from BBAW and BL are cited throughout this document with the shelfmark.

This project has been made possible in part by funding from the Adopt-A-Character program of the Unicode Consortium, and has been supervised by Dr. Deborah Anderson and Rick McGowan.

It was also made possible in part by a grant from the U.S. National Endowment for the Humanities, which funded the Universal Scripts Project (part of the Script Encoding Initiative at UC Berkeley). Any views, findings, conclusions or recommendations expressed in this publication do not necessarily reflect those of the National Endowment of the Humanities.

	10F7	10F8	10F9	10FA
0	𐽰 10F70	𐾀 10F80	▨	▨
1	𐽱 10F71	𐾁 10F81	▨	▨
2	𐽲 10F72	◌𐾂 10F82	▨	▨
3	𐽳 10F73	◌𐾃 10F83	▨	▨
4	𐽴 10F74	◌𐾄 10F84	▨	▨
5	𐽵 10F75	◌𐾅 10F85	▨	▨
6	𐽶 10F76	𐾆 10F86	▨	▨
7	𐽷 10F77	𐾇 10F87	▨	▨
8	𐽸 10F78	𐾈 10F88	▨	▨
9	𐽹 10F79	𐾉 10F89	▨	▨
A	𐽺 10F7A	▨	▨	▨
B	𐽻 10F7B	▨	▨	▨
C	𐽼 10F7C	▨	▨	▨
D	𐽽 10F7D	▨	▨	▨
E	𐽾 10F7E	▨	▨	▨
F	𐽿 10F7F	▨	▨	▨

Letters

10F70	𐽰	OLD UYGHUR LETTER ALEPH
10F71	𐽱	OLD UYGHUR LETTER BETH
10F72	𐽲	OLD UYGHUR LETTER GIMEL-HETH
10F73	𐽳	OLD UYGHUR LETTER WAW
10F74	𐽴	OLD UYGHUR LETTER ZAYIN
10F75	𐽵	OLD UYGHUR LETTER FINAL HETH
10F76	𐽶	OLD UYGHUR LETTER YODH
10F77	𐽷	OLD UYGHUR LETTER KAPH
10F78	𐽸	OLD UYGHUR LETTER LAMEDH
10F79	𐽹	OLD UYGHUR LETTER MEM
10F7A	𐽺	OLD UYGHUR LETTER NUN
10F7B	𐽻	OLD UYGHUR LETTER SAMEKH
10F7C	𐽼	OLD UYGHUR LETTER PE
10F7D	𐽽	OLD UYGHUR LETTER SADHE
10F7E	𐽾	OLD UYGHUR LETTER RESH
10F7F	𐽿	OLD UYGHUR LETTER SHIN
10F80	𐾀	OLD UYGHUR LETTER TAW
10F81	𐾁	OLD UYGHUR LETTER LESH • hooked resh

Combining signs

10F82	◌𐾂	OLD UYGHUR COMBINING DOT ABOVE
10F83	◌𐾃	OLD UYGHUR COMBINING DOT BELOW
10F84	◌𐾄	OLD UYGHUR COMBINING TWO DOTS ABOVE
10F85	◌𐾅	OLD UYGHUR COMBINING TWO DOTS BELOW

Punctuation

10F86	𐾆	OLD UYGHUR PUNCTUATION BAR
10F87	𐾇	OLD UYGHUR PUNCTUATION TWO BARS
10F88	𐾈	OLD UYGHUR PUNCTUATION TWO DOTS
10F89	𐾉	OLD UYGHUR PUNCTUATION FOUR DOTS

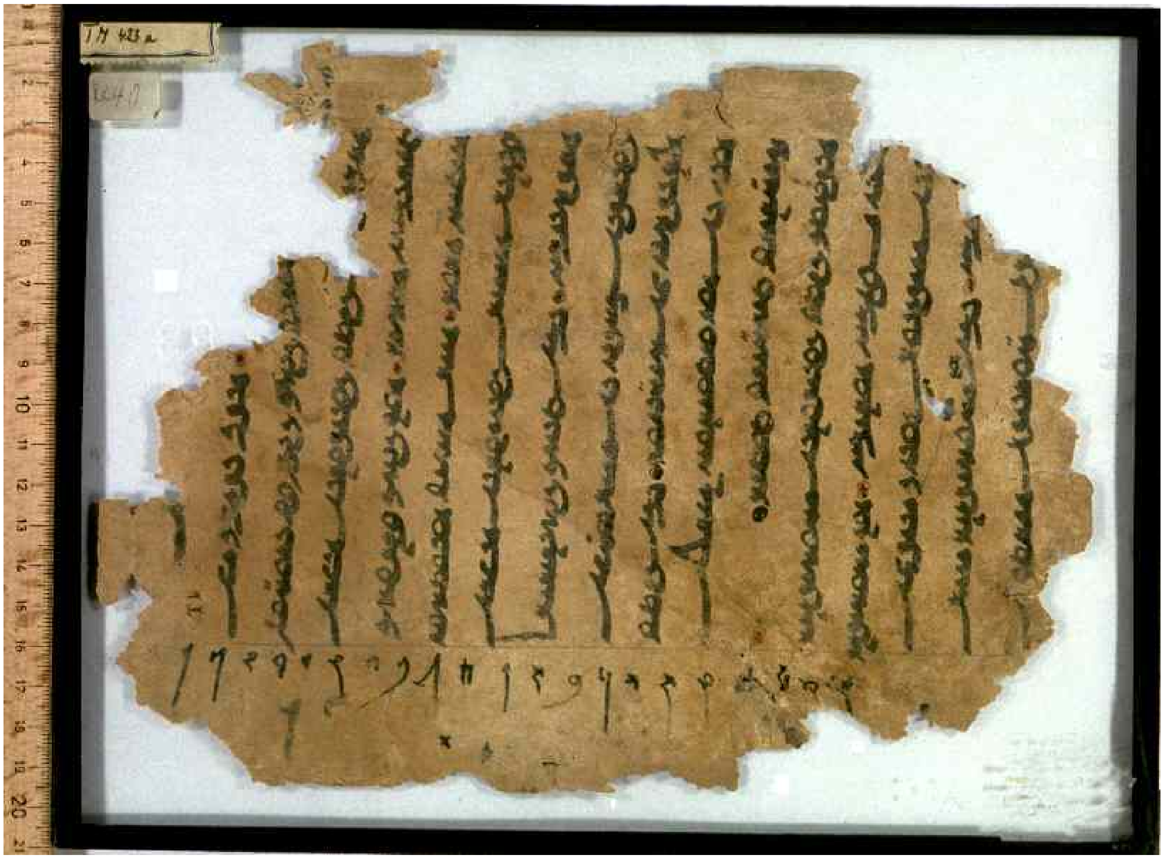
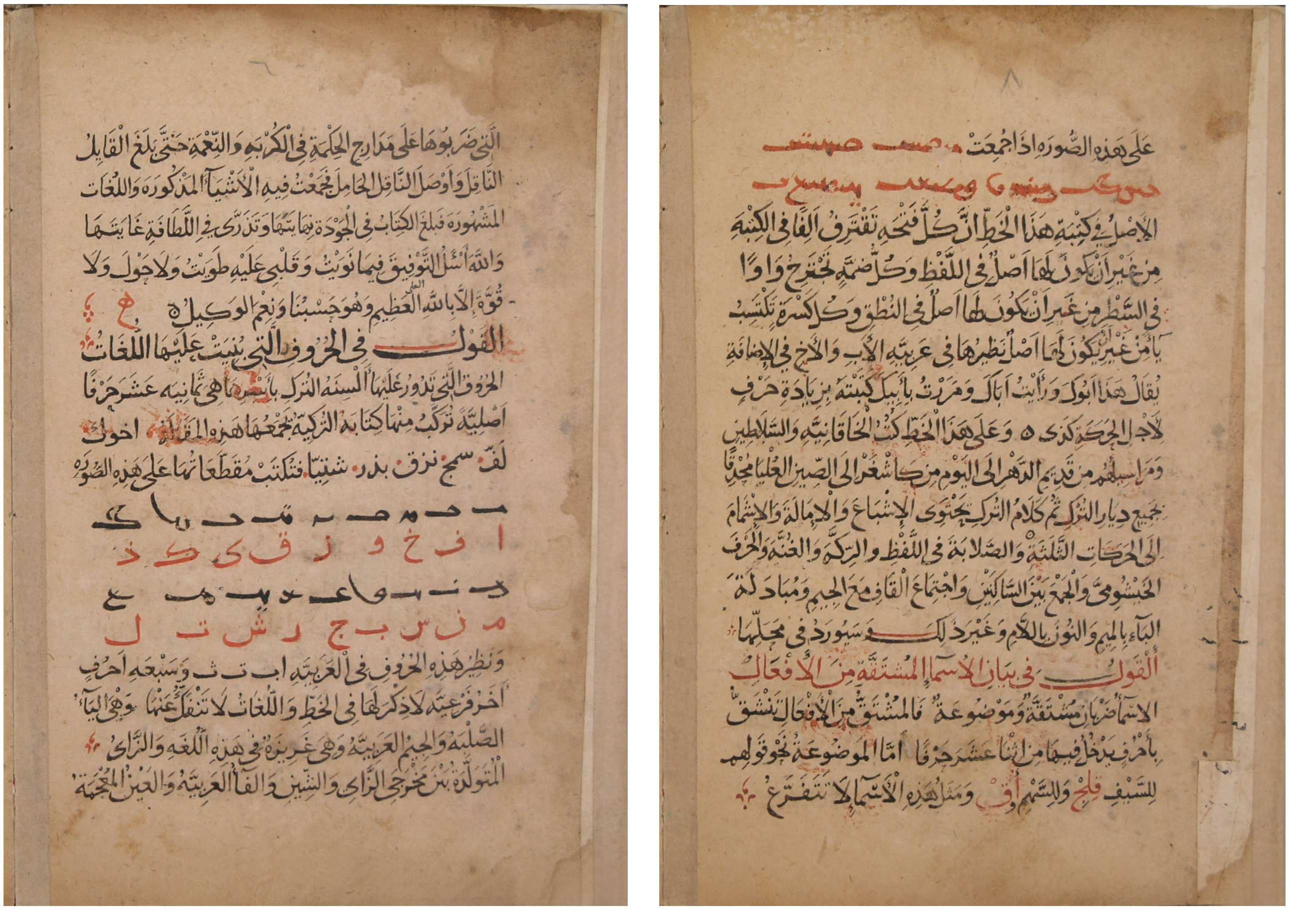


Figure 1: A manuscript from the 9th century (BBAW U 40 recto) with an inventory of Old Uyghur letters in the botom margin (see § 5 for additional details).



التي ضربوها على مدارج الحكمة في الكربة والنعمة حتى بلغ القايل
الناقل واوصل الناقل الحامل فجمعت فيه الاشيا المذكوره واللغات
المشهوره فبلغ الكتاب في الجودة نهايتها وتدرّى في اللطافة غايتها
والله أسأل التوفيق فيما نويت وقلبي عليه طويت ولا حول ولا
قوة إلا بالله العلي العظيم وهو حسبنا ونعم الوكيل ه

القول في الحروف التي بنيت عليها اللغات

الحروف التي تدور عليها ألسنة الترك بأسرها هي ثمانية عشر حرفا
أصلية تركب منها كتابة التركية نجمعها هذه المقالة اخوك
لف سج نزق بذر شنيا فتكتب مقطعاتها على هذه الصوره

ا ف خ و ز ق ى ك ذ

م ن س ب ج ر ش ت ل

ونظير هذه الحروف في العربية ا ب ت ث وسبعة احرف
اخر فرعية لا ذكر لها في الخط واللغات لا تنفك عنها وهي اليا
الصلبة والجيم العربية وهي غريبة في هذه اللغة والزاي
المتولدة بين مخرجي الزاي والشين والفا العربية والعين المعجمة

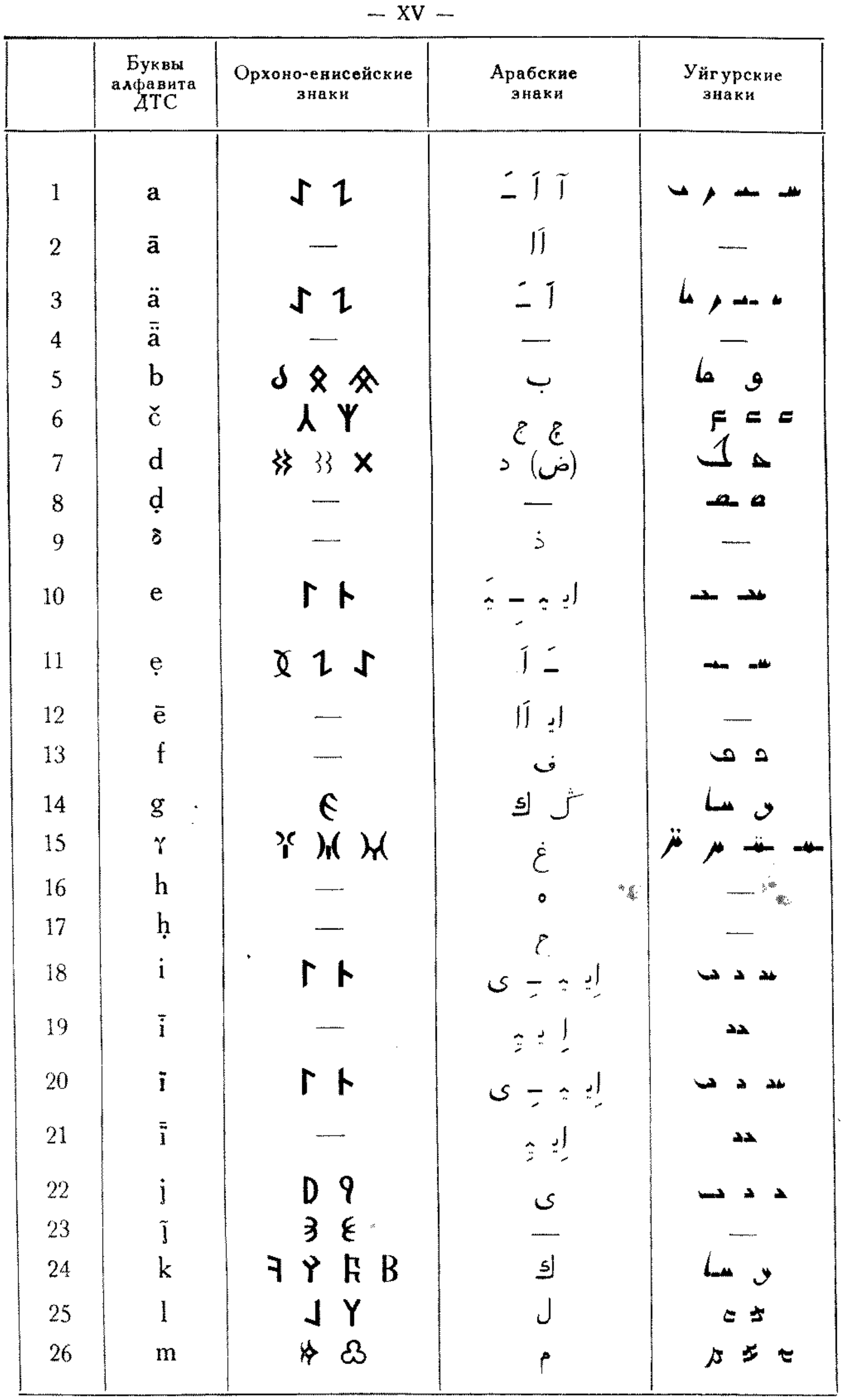
٨

على هذه الصوره اذا اجمعت
الاصل في كتبة هذا الخط ان كل فتحة تفترق الفا في الكتبة
من غير ان يكون لها اصل في اللفظ وكل ضمة تخرج واوا
في السطر من غير ان يكون لها اصل في النطق وكل كسرة تكتسب
يا من غير ان يكون لها اصل نظيرها في عربية الاب والاخ في الاضافة
يقال هذا ابوك ورايت اباك ومررت بابيك كتبته بزيادة حرف
لاجل الحركة كذى ه وعلى هذا الخط كتب الخاقانية والسلاطين
ومراسلهم من قديم الدهر الى اليوم من كاشغر الى الصين العليا محدقا
جميع ديار الترك ثم كلام الترك يحتوي الاشباع والامالة والاشمام
الى الحركات الثلثة والصلابة في اللفظ والركة والغنة والحرف
الخيشومي والجمع بين الساكنين واجتماع القاف مع الجيم ومبادلة
البا بالميم والنون باللام وغير ذلك وسيورد في محلها

القول في بيان الاسما المشتقة من الافعال

الاسما ضربان مشتقة وموضوعة فالمشتق من الافعال تشتق
بأمر فيدخل فيها من اثنا عشر حرفا اما الموضوعة نحو قولهم
للسيف قلج وللسهم اوق ومثل هذه الاسما لا تتفرع

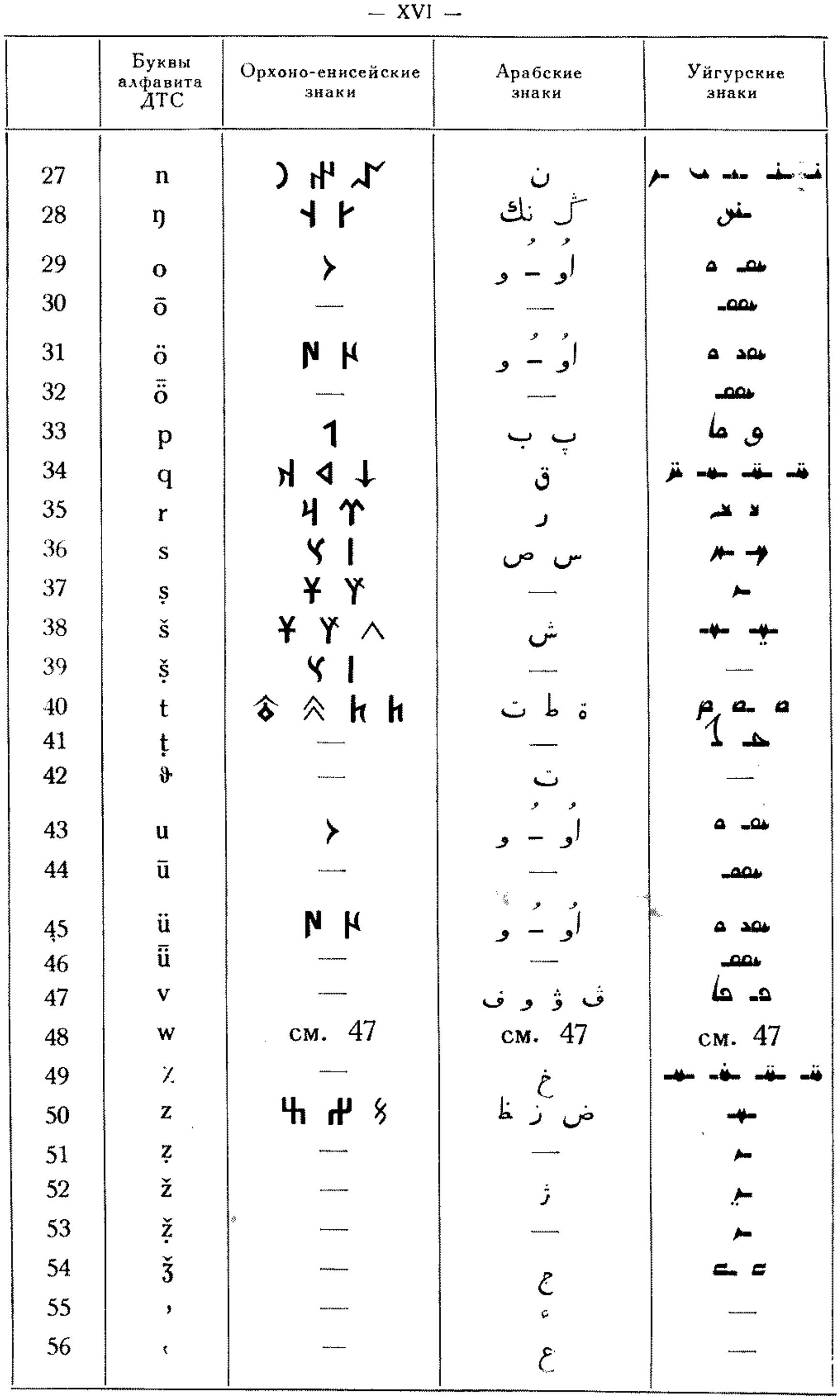
Figure 2: Folios from the *Dīwān luġāt al-turk* by Kāšġarī (11th century). The left folio contains the Old Uyghur repertoire (black ink) with Arabic analogues (red ink). The right folio contains, at top, a mnemonic device with for the Old Uyghur alphabet. This source is significant because it shows Old Uyghur written in a horizontal orientation. Images courtesy of Mehmet Ölmez.



— XV —

	Буквы алфавита ДТС	Орхоно-енисейские знаки	Арабские знаки	Уйгурские знаки
1	a		ـَ أ آ	
2	ā	—	آا	—
3	ä		ـَ أ	
4	ä̅	—	—	—
5	b		ب	
6	č		چ ج	
7	d		د (ض)	
8	ḍ	—	—	
9	δ	—	ذ	—
10	e		ـِ ـ ـِ اِ	
11	ẹ		أ ـَ	
12	ē	—	آا اِ	—
13	f	—	ف	
14	g		ك گ	
15	γ		غ	
16	h	—	ه	—
17	ḥ	—	ح	—
18	i		ی ـِ ـِ اِ	
19	ī	—	ـِ ـِ اِ	
20	ï		ی ـِ ـِ اِ	
21	ï̄	—	ـِ اِ	
22	j		ی	
23	ĵ		—	—
24	k		ك	
25	l		ل	
26	m		م	

Figure 3: Representation of Old Turkic sounds in the Orkhon, Arabic, and Old Uyghur scripts (from Nadeliaev, et al. 1969: xv). Continued in fig. 4.



— XVI —

	Буквы алфавита ДТС	Орхоно-енисейские знаки	Арабские знаки	Уйгурские знаки
27	n	𐰣 𐰤 𐰥	ن	𐽺 𐽺 𐽺 𐽺 𐽺
28	ŋ	𐰭 𐰮	ڭ نك	𐽺𐽷
29	o	𐰆	اُو – وُ	𐽳 𐽳
30	ō	—	—	𐽳𐽳
31	ö	𐰇 𐰈	اُو – وُ	𐽳 𐽳𐽶
32	ȫ	—	—	𐽳𐽳𐽶
33	p	𐰯	پ ب	𐽼 𐽼
34	q	𐰴 𐰶 𐰸	ق	𐽷 𐽷 𐽷 𐽷
35	r	𐰺 𐰼	ر	𐽾 𐽾
36	s	𐰽 𐰾	س ص	𐽻 𐽻
37	ṣ	𐱁 𐱂	—	𐽽
38	š	𐱁 𐱂 𐱃	ش	𐽿 𐽿
39	ş̌	𐰽 𐰾	—	—
40	t	𐱄 𐱅 𐱆 𐱇	ت ط ة	𐾀 𐾀 𐾀
41	ṭ	—	—	𐾀 𐾀
42	ϑ	—	ث	—
43	u	𐰆	اُو – وُ	𐽳 𐽳
44	ū	—	—	𐽳𐽳
45	ü	𐰇 𐰈	اُو – وُ	𐽳 𐽳𐽶
46	ǖ	—	—	𐽳𐽳𐽶
47	v	—	ف ۋ و ڤ	𐽼 𐽼
48	w	см. 47	см. 47	см. 47
49	χ	—	خ	𐽲 𐽲 𐽲 𐽲
50	z	𐰔 𐰕 𐰕	ض ز ظ	𐽴
51	ẓ	—	—	𐽽
52	ž	—	ژ	𐽴
53	ẓ̌	—	—	𐽽
54	ǯ	—	ج	𐽶 𐽶
55	ʼ	—	ء	—
56	ʻ	—	ع	—

Figure 4: Representation of Old Turkic sounds in the Orkhon, Arabic, and Old Uyghur scripts (from Nadeliaev, et al. 1969: xvi). Continued from fig. 3.

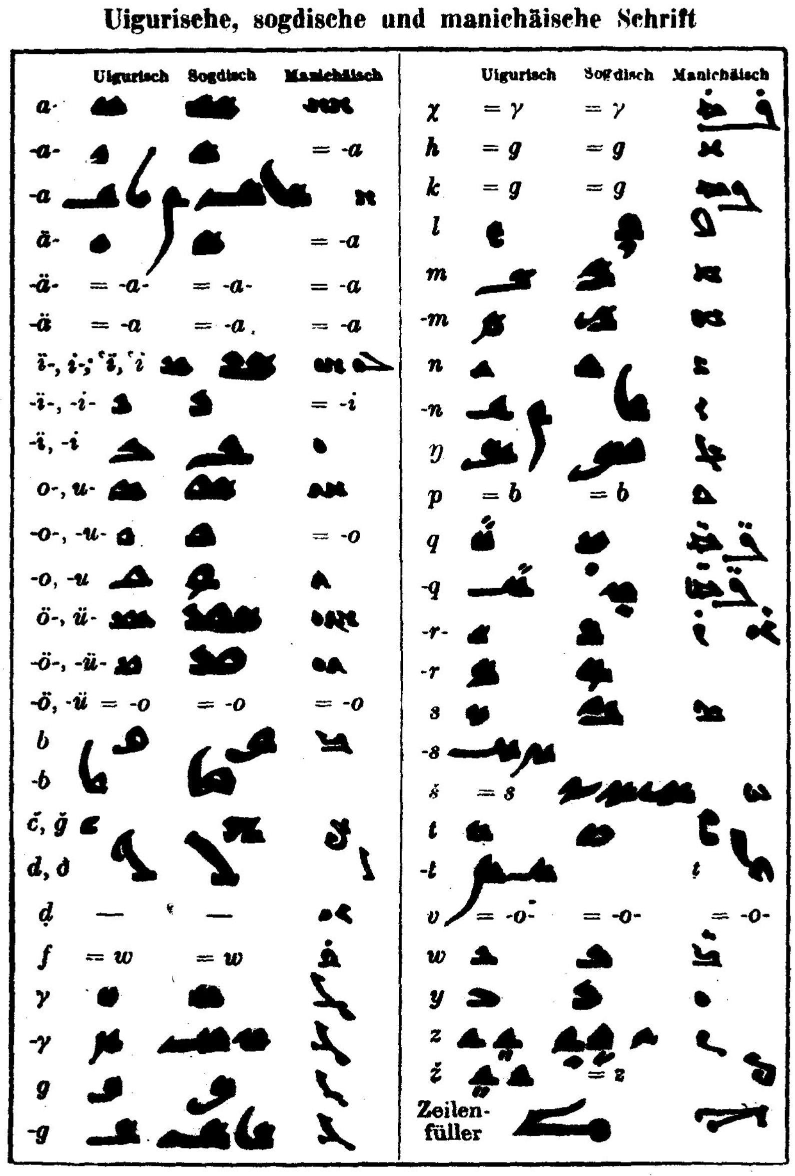

Uigurische, sogdische und manichäische Schrift
Uigurisch Sogdisch Manichäisch
a-
-a- = -a
-a
ä- = -a
-ä- = -a- = -a- = -a
-ä = -a = -a = -a
i-, i-; 'i, 'i
-i-, -i- = -i
-i, -i
o-, u-
-o-, -u- = -o
-o, -u
ö-, ü-
-ö-, -ü-
-ö, -ü = -o = -o = -o
b
-b
č, ğ
d, ð
ḍ — —
f = w = w
γ
-γ
g
-g
Uigurisch Sogdisch Manichäisch
χ = γ = γ
h = g = g
k = g = g
l
m
-m
n
-n
ŋ
p = b = b
q
-q
-r-
-r
s
-s
š = s
t
-t
v = -o- = -o- = -o-
w
y
z
ž = z
Zeilen-füller

Figure 5: Comparison of Old Uyghur, Sogdian, and Manichaean letters (from von Gabain 1950: 17).

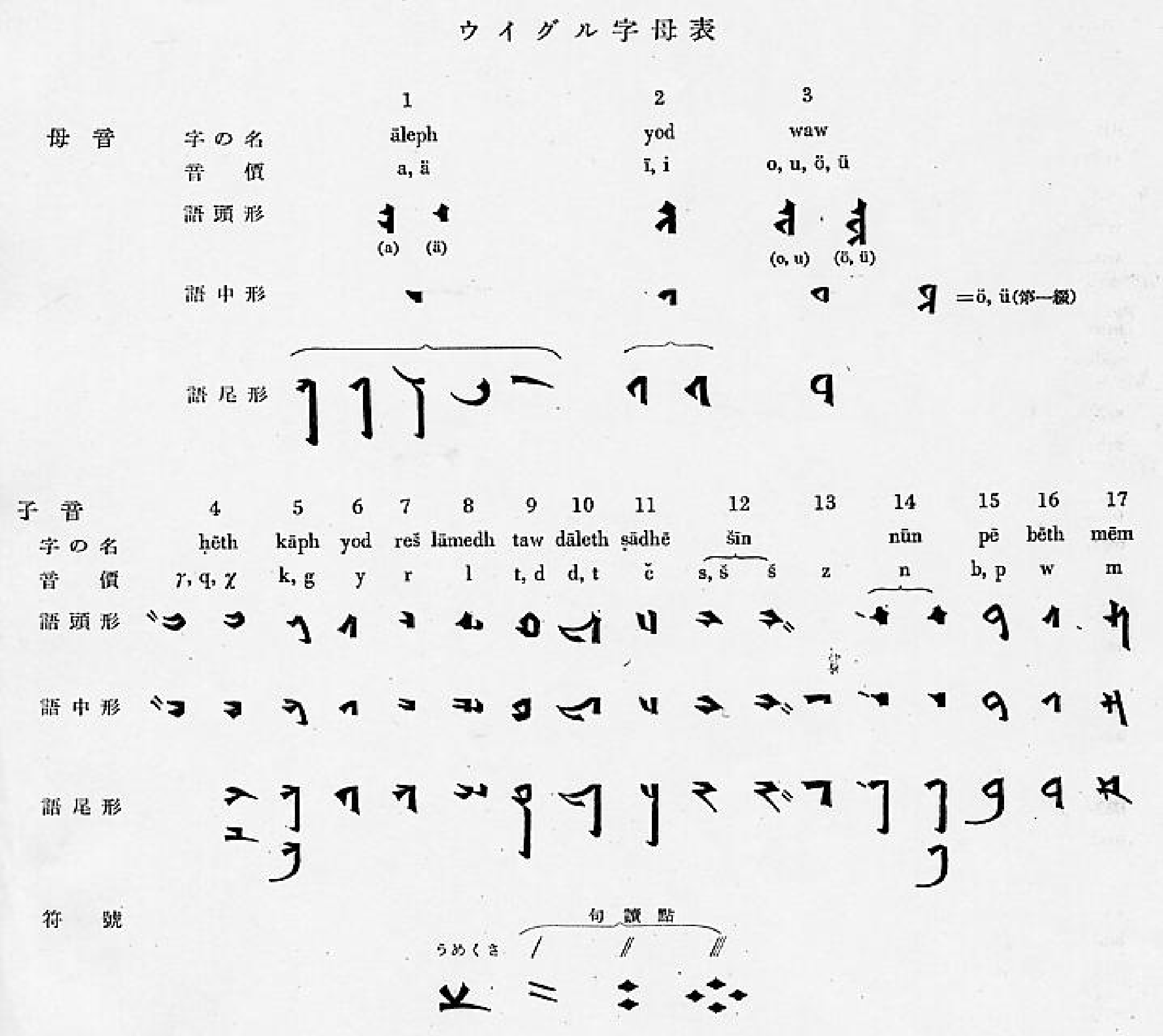
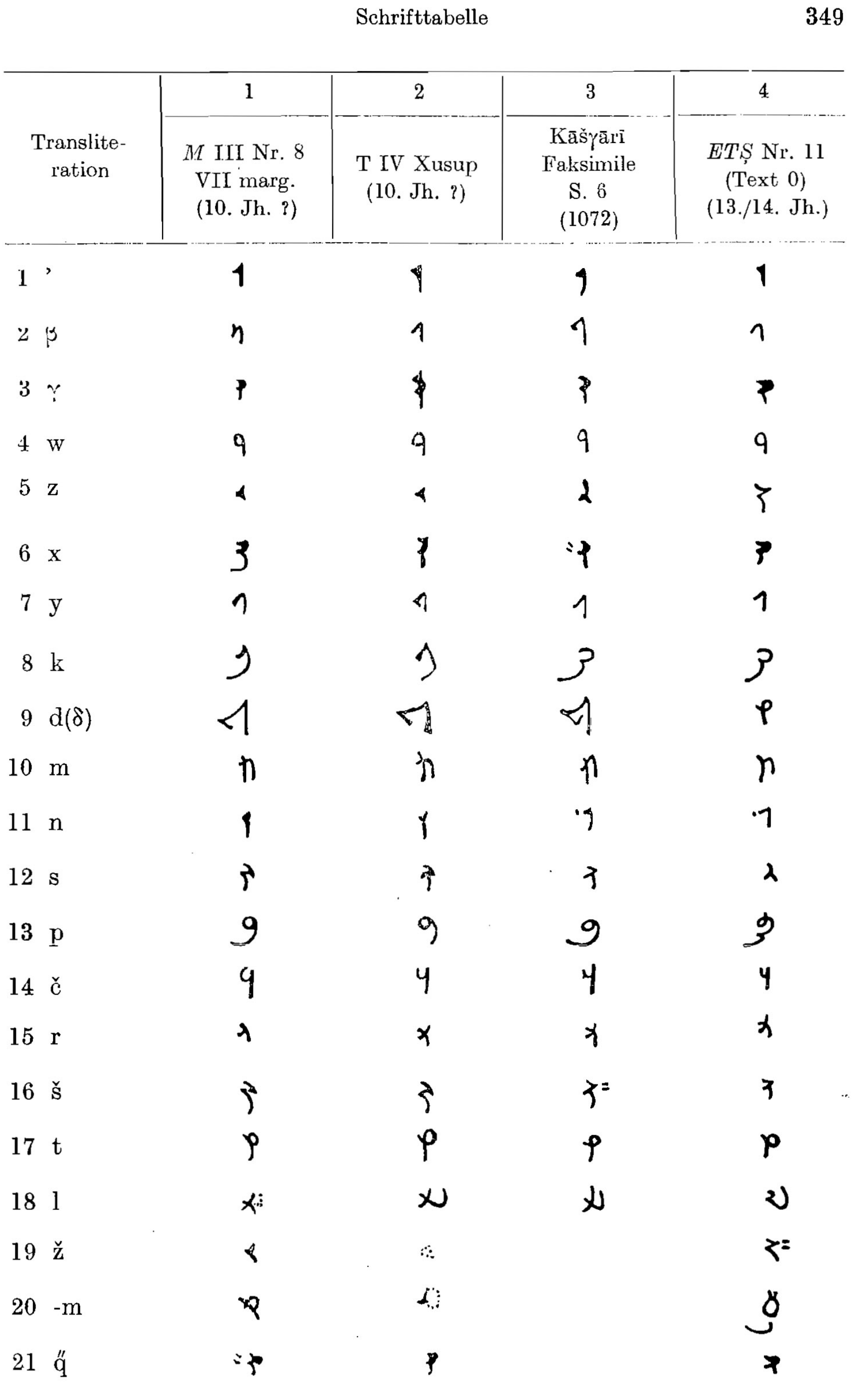

ウイグル字母表
母音
字の名
音価
語頭形
語中形
語尾形
1
2
3
āleph
yod
waw
a, ä
ī, i
o, u, ö, ü
(a) (ä)
(o, u) (ö, ü)
=ö, ü(第一綴)
子音
4 5 6 7 8 9 10 11 12 13 14 15 16 17
hēth kāph yod reš lāmedh taw dāleth ṣādhē šin nūn pē bēth mēm
γ, q, χ k, g y r l t, d d, t č s, š š z n b, p w m
符號
句讀點
うめくさ

Figure 6: Table of Old Uyghur characters used in the Uyghur inscription in the multi-script Yuan dynasty inscriptions at Juyong Guan 居庸關 pass at the Great Wall northwest of Beijing (from Chü-Yung-Kuan 居庸關, “The Buddhist Arch of the Fourteenth Century A.D. at the Pass of the Great Wall Northwest of Peking”, vol. 1, p. 165; reproduced from West 2006). See photograph containing an excerpt of the inscription in fig. 11.

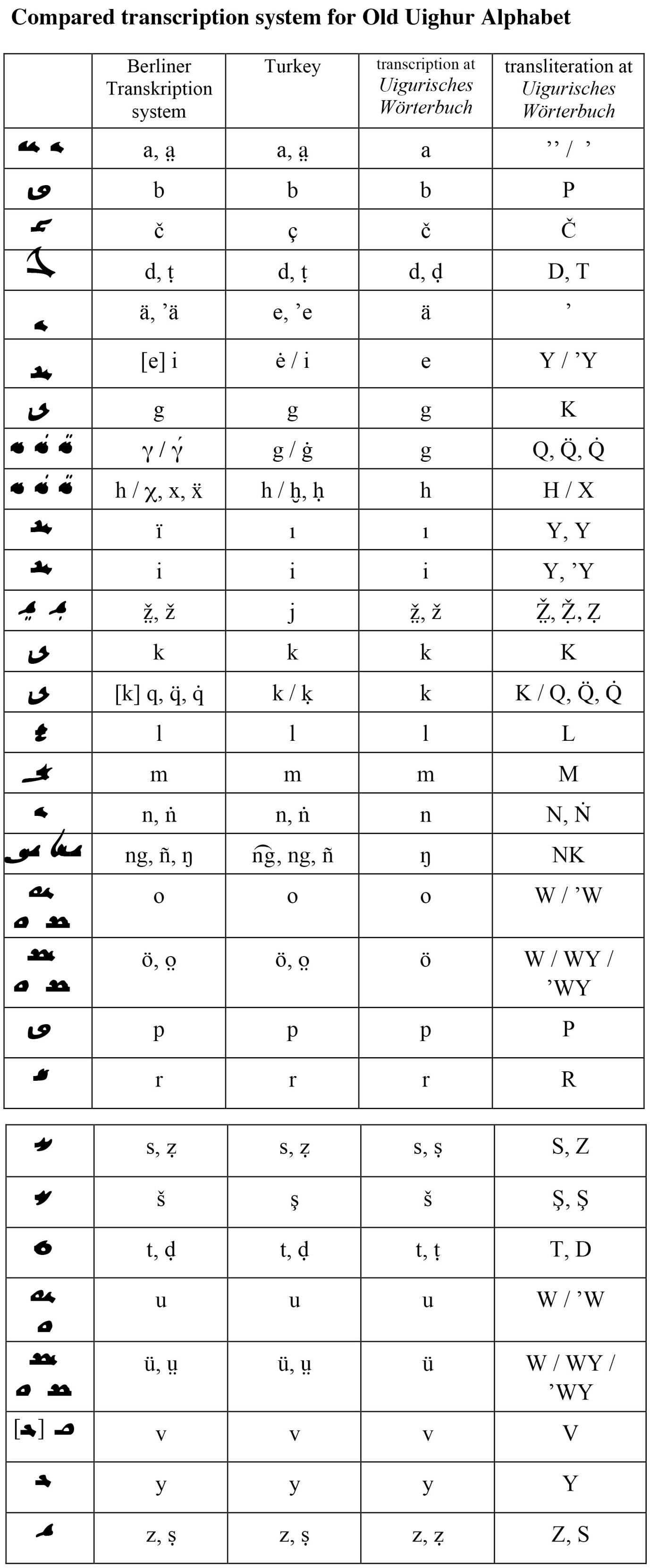
Note: there are a few inaccurate assignment of names for graphemes based upon phonetic value. The glyphs shown for final *beth* (#16) is actually *waw*. The likely reason is that final /b/ does not occur in texts from this period and the original form became obsolete. #13 is unnamed, but it is clearly *zayin*. #10 is not *daleth*, but *lamedh*; *daleth* is not a distinctive letter in Old Uyghur and the name is used in reference to the phoneme /d/. #8 is not *lamedh*, but the ‘hooked’ *resh* (= the LESH proposed here); the name *lamedh* is used as a reference to the phoneme /l/.



Schrifttabelle 349

Transliteration	1 *M* III Nr. 8 VII marg. (10. Jh. ?)	2 T IV Xusup (10. Jh. ?)	3 Kāšγārī Faksimile S. 6 (1072)	4 *ETŞ* Nr. 11 (Text 0) (13./14. Jh.)
1 ’				
2 β				
3 γ̣				
4 w				
5 z				
6 x				
7 y				
8 k				
9 d(δ)				
10 m				
11 n				
12 s				
13 p				
14 č				
15 r				
16 š				
17 t				
18 l				
19 ž				
20 -m				
21 q̋				

Figure 7: Chart showing development and variation in the Old Uyghur script from the 10th through 14th century (from Zieme 1991: 349).



Compared transcription system for Old Uighur Alphabet

	Berliner Transkription system	Turkey	transcription at *Uigurisches Wörterbuch*	transliteration at *Uigurisches Wörterbuch*
	a, a̤	a, a̤	a	’’ / ’
	b	b	b	P
	č	ç	č	Č
	d, ṭ	d, ṭ	d, ḍ	D, T
	ä, ’ä	e, ’e	ä	’
	[e] i	ė / i	e	Y / ’Y
	g	g	g	K
	γ / γ́	g / ġ	g	Q, Q̈, Q̇
	h / χ, x, ẍ	h / ḫ, ḥ	h	H / X
	ï	ı	ı	Y, Y
	i	i	i	Y, ’Y
	ž̤, ž	j	ž̤, ž	Ž̤, Ẓ̌, Ẓ
	k	k	k	K
	[k] q, q̈, q̇	k / ḳ	k	K / Q, Q̈, Q̇
	l	l	l	L
	m	m	m	M
	n, ṅ	n, ṅ	n	N, Ṅ
	ng, ñ, ŋ	n͡g, ng, ñ	ŋ	NK
	o	o	o	W / ’W
	ö, o̤	ö, o̤	ö	W / WY / ’WY
	p	p	p	P
	r	r	r	R
	s, ẓ	s, ẓ	s, ṣ	S, Z
	š	ş	š	Ş, Ş
	t, ḍ	t, ḍ	t, ṭ	T, D
	u	u	u	W / ’W
	ü, ṳ	ü, ṳ	ü	W / WY / ’WY
[ ]	v	v	v	V
	y	y	y	Y
	z, ṣ	z, ṣ	z, ẓ	Z, S

Figure 8: Comparison of transliteration schemes for Old Uyghur (from Ölmez 2016).

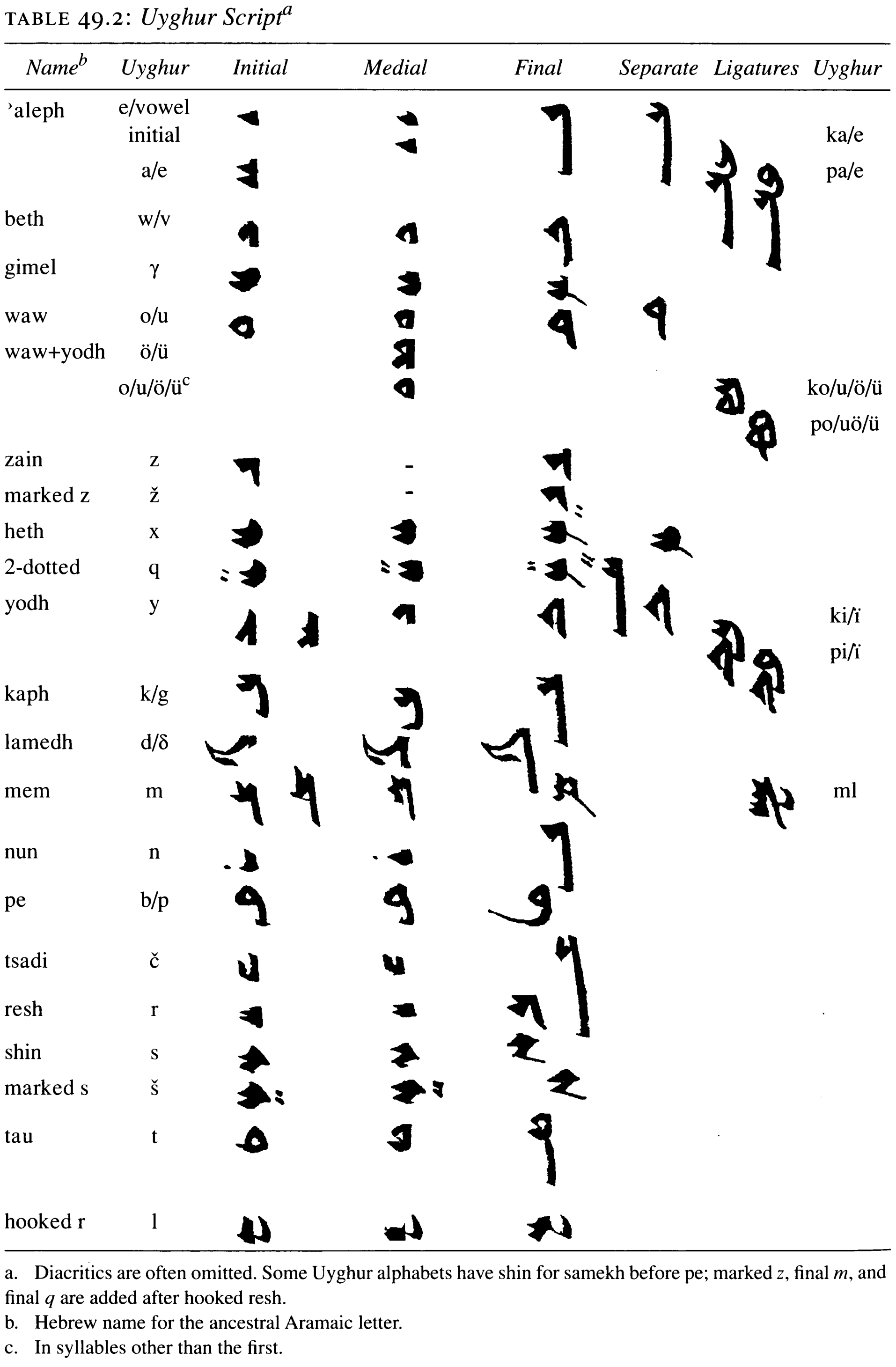


TABLE 49.2: *Uyghur Script*[a]

Name[b]	*Uyghur*	*Initial*	*Medial*	*Final*	*Separate*	*Ligatures*	*Uyghur*
ʾaleph	e/vowel initial						ka/e
	a/e						pa/e
beth	w/v						
gimel	γ						
waw	o/u						
waw+yodh	ö/ü						
	o/u/ö/ü[c]						ko/u/ö/ü
							po/uö/ü
zain	z		-				
marked z	ž		-				
heth	x						
2-dotted	q						
yodh	y						ki/ï
							pi/ï
kaph	k/g						
lamedh	d/δ						
mem	m						ml
nun	n						
pe	b/p						
tsadi	č						
resh	r						
shin	s						
marked s	š						
tau	t						
hooked r	l						

a. Diacritics are often omitted. Some Uyghur alphabets have shin for samekh before pe; marked *z*, final *m*, and final *q* are added after hooked resh.

b. Hebrew name for the ancestral Aramaic letter.

c. In syllables other than the first.

Figure 9: Table showing letters of the Old Uyghur script (from Kara 1996: 540). See table of Mongolian letters from the same source in fig. 10.

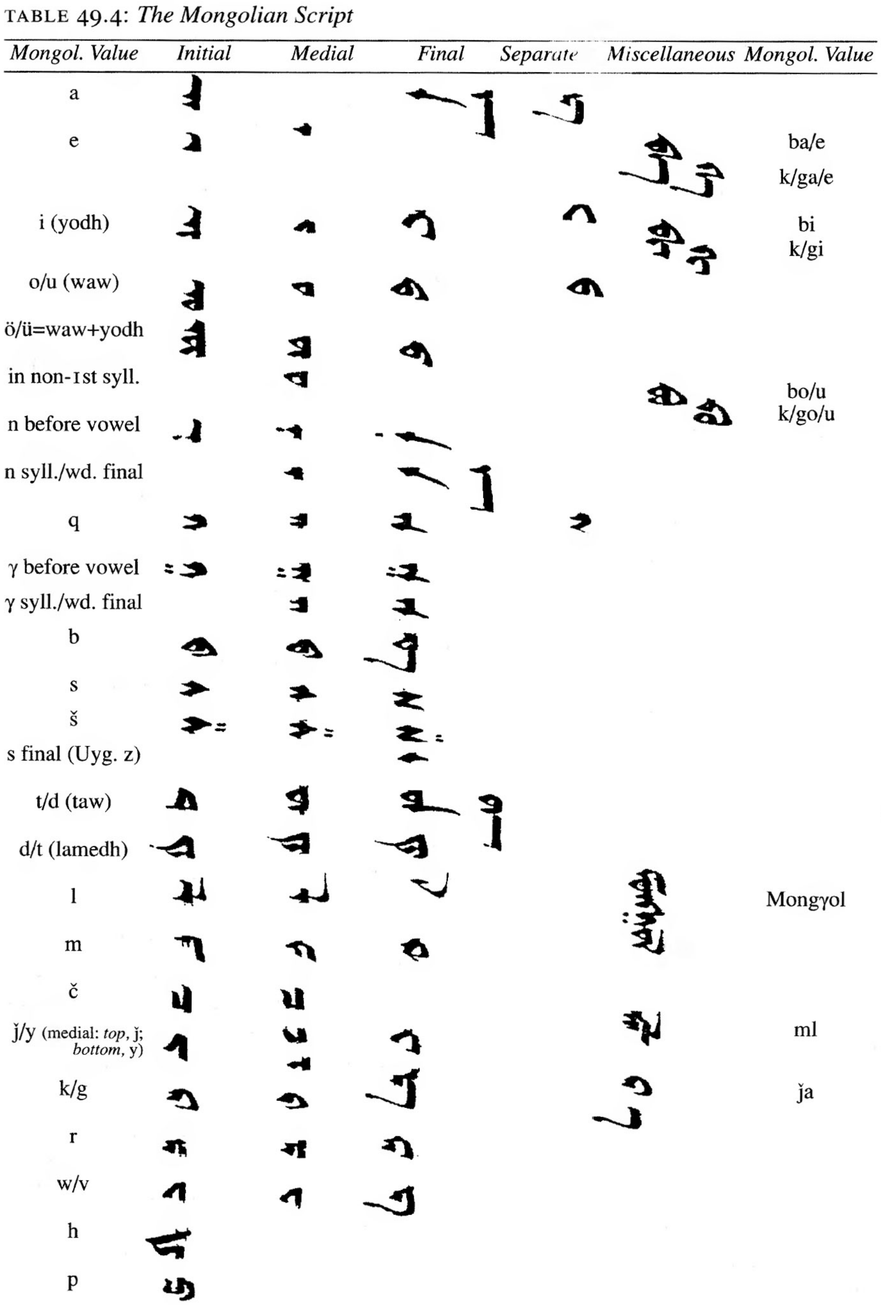


TABLE 49.4: *The Mongolian Script*

Mongol. Value	*Initial*	*Medial*	*Final*	*Separate*	*Miscellaneous*	*Mongol. Value*
a						
e						ba/e k/ga/e
i (yodh)						bi k/gi
o/u (waw)						
ö/ü=waw+yodh						
in non-1st syll.						bo/u k/go/u
n before vowel						
n syll./wd. final						
q						
γ before vowel						
γ syll./wd. final						
b						
s						
š						
s final (Uyg. z)						
t/d (taw)						
d/t (lamedh)						
l						Mongγol
m						
č						
ǰ/y (medial: *top*, ǰ; *bottom*, y)						ml
k/g						ǰa
r						
w/v						
h						
p						

Figure 10: Table showing letters of the Mongolian script (from Kara 1996: 545). See table of Old Uyghur letters from the same source in fig. 9.

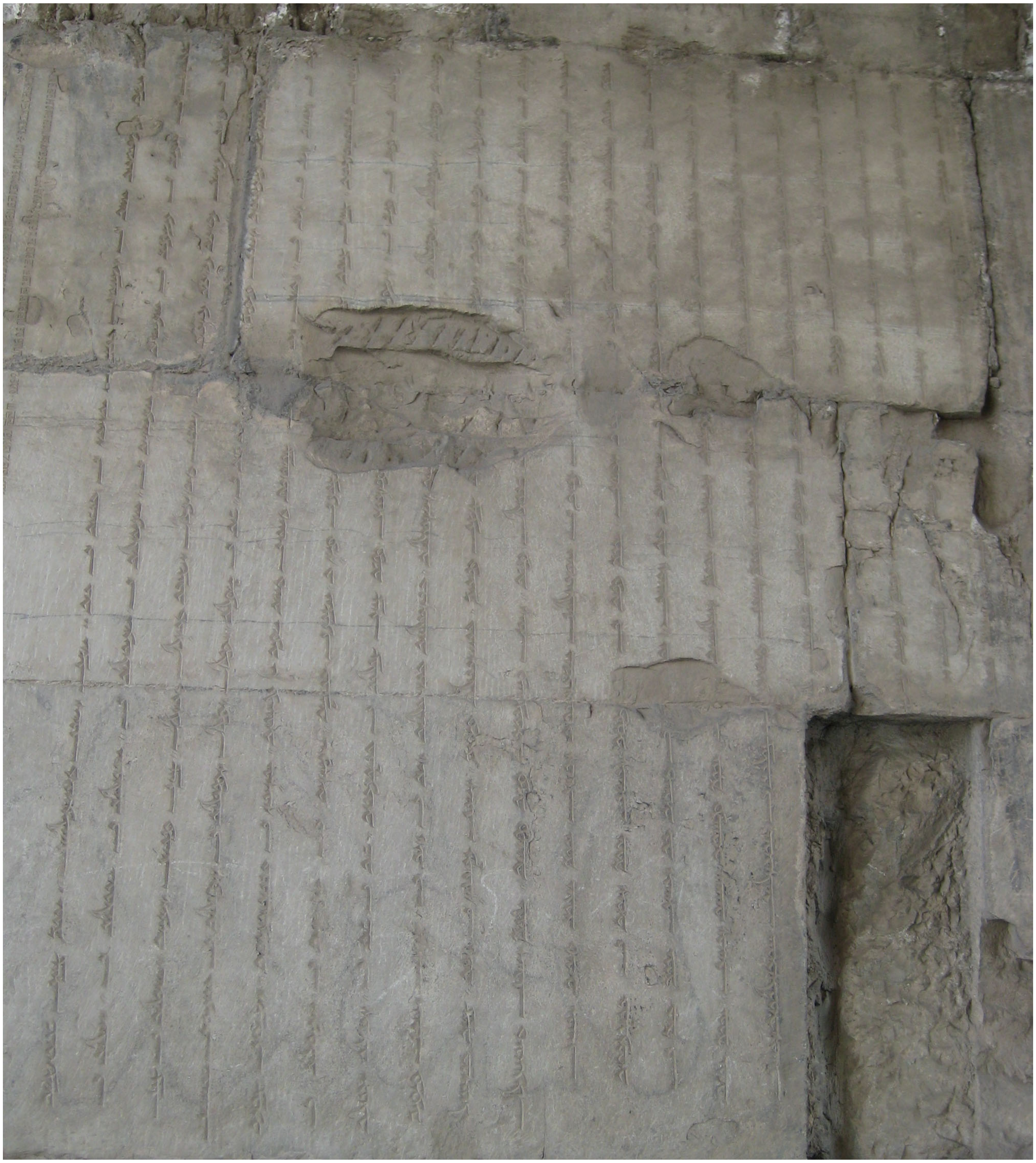


Figure 11: Detail of the Old Uyghur text of the multi-script Yuan dynasty Buddhist inscriptions on the west wall of the Cloud Platform at Juyong Guan 居庸關 pass at the Great Wall northwest of Beijing. Photograph by Andrew West, 2011.

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

A. Administrative

1. **Title:** ***Final proposal to encode Old Uyghur in Unicode***
2. Requester's name: *Anshuman Pandey <pandey@umich.edu>*
3. Requester type (Member body/Liaison/Individual contribution): *Expert contribution*
4. Submission date: *2020-12-18*
5. Requester's reference (if applicable):
6. Choose one of the following:
This is a complete proposal: *Yes*
(or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
a. This proposal is for a new script (set of characters): *Yes*
Proposed name of script: *Old Uyghur*
b. The proposal is for addition of character(s) to an existing block:
Name of the existing block:
2. Number of characters in proposal: *26*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary		B.1-Specialized (small collection)		B.2-Specialized (large collection)	
C-Major extinct	*X*	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?
Anshuman Pandey
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
Anshuman Pandey
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)? *Yes*

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 Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-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, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

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

If YES explain

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

If YES, with whom? *Dr. Dai Matsui <dmatsui@let.osaka-u.ac.jp>*
Dr. Mehmet Ölmez <olmez.mehmet@gmail.com>
Dr. Yukiyo Kasai <yukiyo.kasai@ruhr-uni-bochum.de>
Nicholas Kontovas <n.d.kontovas@hum.leidenuniv.nl>

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: *See text of proposal*

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

Reference: *See text of proposal*

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

If YES, where? Reference: *Currently used by scholars of Turkic and Central Asian studies*

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *N/A*

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? *No*

If YES, is a rationale for its inclusion provided?

If YES, reference:

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, or could be confused with, 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? *Yes*

If YES, is a rationale for such use provided? *Yes*

If YES, reference: *Combining characters for diacritics*

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *N/A*

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 characters? *No*

If YES, are the equivalent corresponding unified ideographic characters identified?

If YES, reference: