Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Review of Preliminary Proposal on the Khitan Large Script (WG2 N5319)**
Source: **Andrew West and Viacheslav Zaytsev**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2025-06-12**

1. Introduction

This document provides a review of the preliminary proposal to encode Khitan Large Script (KLS) submitted by China (WG2 N5319). We welcome this proposal as a step forward in the process of encoding Khitan Large Script, but our initial assessment is that it does not substantially improve on China's previous proposal to encode Khitan Large Script from ten years ago (WG2 N4631), and does not address the points of concern raised by West and Zaytsev in our review of the earlier proposal (WG2 N4559). Indeed, the new proposal provides less information about individual KLS characters than the old proposal did. The new proposal omits the source references for each evidence glyph that were provided in the old proposal, so we are unable to use the new proposal to check the proposed characters in their original source. The new proposal also omits the input codes which the old proposal provided, which were very useful for finding individual characters in the proposal, and verifying the sort order of the proposed characters.

A mapping table between the entries in the old and new proposals is unfortunately not provided. This information is essential, as it would enable us to update our Excel spreadsheet of the proposed KLS characters in N4631 with the new sequence numbers in N5319, and thereby allow us to mitigate some of the deficiencies in the new proposal.

2. Glyph Encoding versus Character Encoding

There are no surviving dictionaries or linguistic texts that list all KLS characters, and no printed texts for KLS which would allow us to determine the orthodox glyph forms of characters. Instead, we have to rely on a limited corpus of epigraphical material (about twenty epitaph inscriptions on stone tablets, one complete stele inscription, about a dozen stele fragments, fragments of ink inscriptions on the inner walls of tombs, a couple of rock inscriptions, and some inscriptions on coins, charms, seals of office, and other artefacts). The main source for KLS characters are epitaph inscriptions, which are in variable states of

preservation and legibility, and can only be partially deciphered. This makes the task of identifying the repertoire of KLS characters, and their correct glyph forms extremely difficult.

The proposals from China list all the glyph forms that can be identified from all the available sources, but do not attempt to unify variant glyph forms that represent the same abstract character. Thus, the proposals are essentially requesting a glyph encoding for Khitan Large Script. We believe that this goes against the fundamental principles of the Universal Coded Character Set (UCS), which encodes characters not glyphs.

We think that compiling a list of glyph forms is a useful first step in preparing an encoding proposal, but such a list cannot be considered to be a character repertoire suitable for encoding in the UCS. The next steps are to determine: a) what abstract character each glyph form corresponds to; and b) what is the most suitable representative glyph form to use for each abstract character. This is long, time-consuming, and difficult task, but it is essential. In this respect, provision of a single cut image of a character from an inscription is not very helpful, as what is really required is the context in which the character occurs, so we can determine whether it occurs in the same position in the same collocation as other graphically-similar characters (which would indicate that the characters are glyph variants of the same abstract character).

3. Case Study

In our previous review document (N4559) we noted as an example that 没 (N4631 #1791; N5319 #1765) and 没 (N4631 #1797; N5319 #1771) are glyph variants of the same abstract character as they occur interchangeably in the common Khitan word 未卅没/未卅没 meaning ‘preface’ or ‘beginning’. There are very many entries in N5319 which we suspect are unifiable glyph variants, for example nos. 47 through 50 are likely to all be forms of the same common character meaning ‘time’.

Here we provide a case study of one particular character, where textual analysis can demonstrate that at least seven entries listed in N5319 are glyph variants of the same abstract character, and should be unified for encoding purposes. Table 1 lists seven entries from N5319 that we think are simple glyph variants of the same abstract character, with minor differences in the realization of glyph form due to the degree in which the character is written in a semi-cursive script.

Table 1: Entries in N5319 for Glyph Variants of the Same Abstract Character

No.	Code	Font Glyph	Source Image	Tables 2–3 Example
673	FC596	歩		11

No.	Code	Font Glyph	Source Image	Tables 2–3 Example
702	FC5B3	岁		2
945	FC6A6	岁		
958	FC6B3	岁		6
959	FC6B4	岁		2
987	FC6D0	岁		
2141	FCB52	岁		

* NB the same source image is given for nos. 702 and 959, with almost identical font glyphs; and the same source image is given for nos. 987 and 2141, with the font glyphs differing by the presence or absence of a dot stroke at the top right.

The abstract character that these glyph variants represent commonly occurs following the word 仁仅 (identified as meaning the court position of *zhīhòu* 祗候 'Usher') in two collocations, for which we provide eleven examples from five epitaph inscriptions and one incomplete example from the tomb wall inscriptions of Yelü Pugu 耶律蒲古 (d. 1031):

- 仁仅岁岑 (Examples 1–8)
- 仁仅岁凭 (Examples 9–11)
- 仁仅岁□ (Example 12)

We can see from the twelve examples shown in Table 2 and Table 3 that the character 岁 occurs in various slightly different glyph forms, including nos. 673, 702/959, and 958, but it is clear that all examples reflect the same abstract character. It makes no sense to encode all glyph forms of the same abstract character as separate characters in the UCS. Therefore, the seven entries listed in Table 1 should be unified, and a single representative glyph form selected (we recommend no. 673 as this appears to be the regular script form).

Table 2: Examples 1–6

1	2	3	4	5	6
					
Dorlipun line 9	Yelü Changyun line 11	Xiao Paolu lines 2–3	Xiao Paolu line 5	Yelü Qi line 10	Yelü Qi line 12

Table 3: Examples 7–12

7	8	9	10	11	12
					
Yelü Qi line 15	Yelü Qi line 16	Yelü Qi line 35	Yelü Xinie line 8	Yelü Xinie line 14	Yelü Pugu

The exercise we carried out for this one KLS character needs to be performed, as far as is possible, for all 2,209 entries listed in N5319. We understand that this would be a huge, multi-year project, but it needs to be done, and when it has been completed it will have greatly improved our understanding of Khitan Large Script.

4. Discrepancy in Character Count and Character Table Contents

Section 5 ("A Plan for Computer Processing of Khitan Large Script") of N5319 claims that 2,245 characters were identified after screening inscriptions and excluding duplicates. This appears as a notable increase from the 2,218 characters listed in the table of the previous proposal (N4631), which would be a welcome development. However, the table accompanying N5319 (Section 7) contains only 2,209 characters, nine fewer than in N4631. This contradiction between the text's claim and the table's contents raises serious concerns about the proposal's accuracy and reliability.

The proposal provides no explanation for why the table lists fewer characters than both the 2,245 stated in the text and the 2,218 in N4631. Possible reasons could include the exclusion of duplicate glyphs, removal of erroneous forms, or an incomplete table, but the absence of clarification leaves these questions unanswered. The text mentions that Professor Sun Xichun was entrusted to "rewrite the forms of the large characters of the Khitan, correct some incorrect writing," and "add some characters that were previously missing," yet it fails to specify which characters were added, removed, or corrected to account for the claimed 2,245 or the table's 2,209. Combined with the lack of a mapping table between N4631 and N5319, this makes it impossible to verify the nature of the revisions or understand the reduction in the table's character count.

5. Evidence of a New Font but Lack of Transparency in Glyph Corrections

The new proposal employs a newly developed Khitan Large Script font, visible in the document, which corroborates the claim in Section 1 that Professor Sun Xichun was tasked with rewriting the forms of Khitan Large characters. The presence of this font indicates that serious work was undertaken to redraw glyphs, likely including the correction of "some incorrect writing" as stated. This tangible output suggests efforts to improve the representation of Khitan Large Script characters, potentially addressing errors in glyph forms from N4631.

However, the proposal lacks specific details about which glyphs were corrected or how the new font differs from previous representations. It does not provide before-and-after comparisons of corrected glyphs, nor does it identify the inscriptions from which errors were identified and fixed. Similarly, the claim of adding "previously missing" characters is not supported by examples or references to their sources. This absence of transparency hinders verification of the corrections' scope and significance. Without such evidence, it remains unclear whether the new font reflects substantial improvements or minor adjustments to existing glyph forms. Providing specific examples of corrected or newly added glyphs would have enhanced the proposal's credibility and utility.

6. Potential Superficiality and Overlap with Previous Work

The extensive reuse of content from N4631, particularly in Sections 2 (“Creation and Application of Khitan Large Script”), 3 (“Existing Materials of Khitan Large Script”), and 4 (“Nature of Khitan Large and Small Script”), which are nearly identical to their counterparts in N4631, is understandable, as authors may retain well-formulated context. However, this indicates limited new research, especially considering over a decade has passed since 2014. The new font and Section 5’s plan for computer processing (e.g., stroke-based classification, input method, font development) demonstrate some progress, but the proposal’s lack of concrete evidence for claimed revisions raises concerns about its superficiality.

Section 5 describes processes such as creating character cards, hiring calligraphers, and developing an input method, but these are presented theoretically, with the new font being the only visible output. The proposal does not clarify whether these efforts represent new work or formalize activities already underway during N4631’s preparation. The discrepancy between the claimed 2,245 characters and the table’s 2,209, coupled with the reduction from N4631’s 2,218, further suggests that the revisions may be minimal or even regressive. Without a detailed changelog or mapping table, as highlighted in Section 1 of this document, it is difficult to assess whether N5319 substantively advances N4631 or merely repackages existing work with minor updates.

7. Issues with Source Images

The new proposal exhibits significant issues with source images, repeating problems we identified in our 2014 review (N4559) of the previous KLS proposal (N4631). Specifically, multiple cases exist where different characters (characters assigned distinct PUA codepoints in the proposal) share the same source image, indicating errors in the repertoire. Additionally, at least one character lacks a source image entirely, and many source images are inadequate for encoding purposes.

Due to the short timeframe available for this review, we have not had the opportunity to examine all source images in N5319. Nevertheless, we have identified the following pairs of characters that share identical source images, with varying interpretations of the resulting font glyphs, in addition to the two pairs already mentioned in Section 3 of this document (nos. 702 and 959, 987 and 2141):

> Nos. 174 and 175: Identical source image, same font glyph (with different PUA codepoints).
>
> Nos. 583 and 1590: Identical source image, same font glyph (with different PUA codepoints).
>
> Nos. 639 and 640: Identical source image, different font glyphs; the source image for no. 639 erroneously repeats the image for no. 637, with the intended image for no. 639 incorrectly placed beneath it, showing only the top stroke.

Nos. 887 and 2194: Identical source image, different font glyphs.

Nos. 909 and 2202: Identical source image, slightly different font glyphs.

Nos. 990 and 1503: Identical source image, slightly different font glyphs.

Nos. 1344 and 1388: Identical source image, different font glyphs.

Nos. 1781 and 1783: Identical source image, slightly different font glyphs.

Furthermore, character no. 1964 lacks a source image, making it impossible to verify its origin. These findings echo our critique in N4559, where we listed 30 cases in N4631 where different characters shared the same source image and noted missing source images for several characters.

Another major concern we have is that a large number of entries in N5319 do not show a source image from a rubbing or photograph of an inscription, but only provide the image of a hand-written glyph form (e.g., nos. 13–15, etc.). We do not consider that this is acceptable evidence for encoding. For the Khitan Small Script (KSS), which has been extensively studied by numerous scholars, we used tables of characters given in various secondary sources as evidence for encoding individual characters (see WG2 N4725R Table 5), but there are no similar secondary sources for Khitan Large Script, so it is necessary to support each proposed KLS character with images of rubbings or photographs of original inscriptions. Moreover, each source image should be accompanied by a reference to the source inscription, so that it can be verified. The absence of such references and the use of hand-drawn glyphs significantly undermine the proposal's credibility. This issue was also raised in N4559, where we emphasized the need for rubbings or photographs to ensure glyph accuracy.

In conclusion, the persistence of these issues in N5319, despite our prior feedback, is regrettable, highlighting the need for enhanced focus on the essential aspects of source documentation and glyph fidelity, which are vital for a robust encoding proposal.

8. Recommendations for Improvement

To address the deficiencies identified in this review and unresolved issues from our prior feedback, we recommend the following for future revisions of the Khitan Large Script encoding proposal:

Resolve the character count discrepancy: Clarify why the table lists 2,209 characters while the text mentions 2,245, and how this compares to the 2,218 characters in N4631. Provide a changelog or mapping table to document additions, removals, or corrections.

Substantiate glyph corrections: Include specific examples of corrected glyphs (e.g., before-and-after images), deleted and newly added characters, with references to their source inscriptions, to validate the improvements reflected in the new font.

Restore omitted information: Reinstate source references and input codes from N4631, or provide equivalent mechanisms to trace characters to their origins and facilitate verification.

Prioritize character unification: As emphasized in Section 3, focus on unifying glyph variants into abstract characters using textual analysis, aligning with UCS principles. Provide occurrence counts for each character in the KLS corpus to distinguish consistent forms from errors or one-off variants, as suggested in our 2014 review (N4559).

Address CJK unification: Analyze the overlap between KLS characters and CJK unified ideographs, as approximately 18% of N4631's repertoire was found to be identical or similar to CJK characters (see our N4559). Provide a mapping of such correspondences, address potential security risks of encoding KLS clones, and seek guidance from the Unicode Technical Committee (UTC) on whether unification with CJK ideographs is appropriate.

Ensure accurate source images: Use rubbings or photographs of original sources as evidence images, accompanied by specific source references, rather than hand-drawn glyphs, which are prone to errors. Resolve cases where different characters share the same source image and address missing source images, as detailed in Section 7 of this document, to ensure all characters have unique, verifiable source images and references, avoiding discrepancies between proposed glyphs and their original forms.

Acknowledge and address prior feedback: Recognize and respond to concerns raised in our 2014 review (N4559), which is notably absent from N5319's bibliography. None of the issues highlighted in N4559— such as the need for glyph unification, CJK overlap analysis, accurate source images, etc.—have been addressed in N5319. Future revisions should explicitly cite and engage with prior feedback to demonstrate progress and ensure alignment with community expectations.

These steps would enhance the proposal's transparency, rigor, and alignment with UCS standards, addressing both technical shortcomings and concerns about its substantive contributions raised in this review and our 2014 feedback (N4559). As scholars deeply engaged in Khitan studies, we fully appreciate the complexities involved in encoding Khitan Large Script and establishing its repertoire. We do not expect all the recommendations outlined above to be fully implemented, as the challenges inherent in this process are significant and may render complete resolution taking years. However, fundamental requirements, such as providing a mapping table between N4631 and N5319 and including clear source references for glyphs, are essential for any encoding proposal and should have been addressed inherently, particularly given that source references were provided in N4631, albeit in an "encoded" form. Nevertheless, we remain open to collaboration and are committed to supporting efforts to advance the encoding of Khitan Large Script in a rigorous and transparent manner.

9. An Alternative Way Forward

No real progress has been made in encoding the Khitan Large Script over the last ten years, and we feel that perhaps one group of experts submitting monolithic proposals every few years is not an appropriate or feasible way to achieve the encoding of a script with such a large number of undeciphered characters. An alternative solution that could be considered is to establish an online collaborative platform where individual Khitan experts from all countries could propose and review individual characters. After a reasonable number of characters have been reviewed and accepted by experts, then a joint proposal could be submitted for their encoding. The platform would continue to remain active for experts to propose and discuss additional characters for inclusion in future proposals for Khitan Large Script extensions.

The KLS online platform could be modelled on the Online Review Tool (ORT) used with great success by the Ideographic Research Group (IRG) to review CJK characters for encoding, and could be hosted by the Unicode Consortium. For each proposed character there would be a dedicated page which shows images of the character in context (not just a cut image of the glyph in isolation) from all sources where the character occurs. Registered experts would be able to make comments on the page, provide additional evidence, and suggest unifications.

10. Concluding Remarks

Despite some critical remarks above, we sincerely thank our colleagues and the authors of N5319 for their continued contributions to the challenging endeavour of encoding the Khitan Large Script.

We very much welcome support and advice from Khitan experts in China and elsewhere in order to advance a collaborative approach to encoding Khitan Large Script. The successful Khitan Small Script proposal was made possible through the help and financial support of the Script Encoding Initiative who coordinated an international meeting in 2016, thereby facilitating broad expert participation, and demonstrating the value of collective efforts in such complex projects. This engagement and support overcame a six-year stalemate, enabling the encoding of KSS to progress rapidly and be included in the Unicode and ISO/IEC 10646 standards in 2020. Regrettably, the encoding of Khitan Large Script presents far greater challenges than anticipated, and without significant efforts from all stakeholders to overcome the obstacles, there is a high risk that outstanding issues will remain unresolved, and Khitan Large Script will not be encoded in the foreseeable future.

11. References

N4559. *Andrew West*, *Viacheslav Zaytsev*. Preliminary Review of Proposal on Encoding Khitan Large Script in UCS (N4631). JTC1/SC2/WG2 N4559 = L2/14-233. 2014-10-14.

URL: https://www.unicode.org/wg2/docs/n4559.pdf
URL: https://www.unicode.org/L2/L2014/14233-n4559.pdf

N4631. *China*. Proposal on Encoding Khitan Large Script in UCS. ISO/IEC JTC1/SC2/WG2 N4631 = L2/14-234 / Project Director: Wu Yingzhe, Sun Bojun, Nie Hongyin; Proposal Writing: Wu Yingzhe; Draft Writing: Dou Minli; Producer of Khitan Large Character: Huaguang Group; Annotation of the Khitan Small Script: Wu Yingzhe, Jiruhe, Hugejileetu. 2014-09-23.

URL: https://www.unicode.org/wg2/docs/n4631.pdf
URL: https://www.unicode.org/wg2/docs/n4631%201of5.pdf
URL: https://www.unicode.org/wg2/docs/n4631%202of5.pdf
URL: https://www.unicode.org/wg2/docs/n4631%203of5.pdf
URL: https://www.unicode.org/wg2/docs/n4631%204of5.pdf
URL: https://www.unicode.org/wg2/docs/n4631%205of5.pdf
URL: https://www.unicode.org/L2/L2014/14234-n4631-khitan.pdf

N4725R. *Andrew West*, *Viacheslav Zaytsev*, *Michael Everson*. Towards an Encoding of the Khitan Small Script. JTC1/SC2/WG2 N4725R = L2/16-113. 2016-05-21.

URL: https://www.unicode.org/wg2/docs/n4725r-khitan-small-script.pdf
URL: https://www.unicode.org/L2/L2016/16113r-n4725r-khitan-small-script.pdf

N5319. *China*. A Preliminary Proposal on the Khitan Large Script and to be a substitute for the previous one (N4631). ISO/IEC JTC1/SC2/WG2 N5319 / Authors: Project Director: Sun Bojun, Wu Yingzhe, Nie Hongyin, Wuyu; Draft Writing: Sun Xichun; Producer of Khitan Large Character: Zhonghua Ziku, Huaguang Group; Annotation of the Khitan Large Script: Wu Yingzhe, Jiruhe, Hugejileetu. 2025-06-06.

URL: https://www.unicode.org/wg2/docs/n5319-KhitanLargeScriptEncoding.pdf