L2/15-342
2015-12-30

Proposal to encode a SUBJOINER for Zanabazar Square

Anshuman Pandey
Department of Linguistics
University of California, Berkeley
Berkeley, California, U.S.A.
anshuman.pandey@berkeley.edu

December 30, 2015

1 Introduction

This proposal has three objectives:

1. Encode the new character ◌ZANABAZAR SQUARE SUBJOINER, with properties as specified in §3.
2. Transfer the conjunct-forming function of ◌ZANABAZAR SQUARE SIGN VIRAMA to the SUBJOINER.
3. Redefine VIRAMA to function solely as a silencer of the inherent vowel.

These recommendations do not alter the encoding model for the Zanabazar Square script, which has already been approved for inclusion in Unicode (see L2/14-024 and the updated proposal L2/15-337). Rather, they improve and simplify the model by distributing two distinct functions currently embodied within a single character across two separate characters dedicated to one function each.

2 Justification

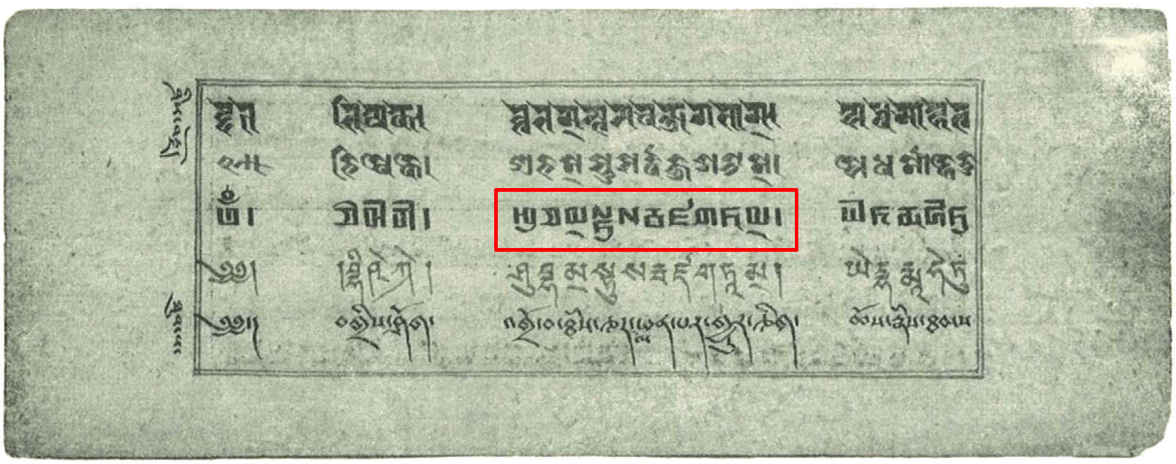
The traditional repertoire for Zanabazar Square does not have a native *virāma* or *halanta*. The ◌ZANABAZAR SQUARE FINAL CONSONANT MARK is conceptually similar, but is used only for marking a bare consonant in syllable-final position in Mongolian and Tibetan linguistic contexts. Bare consonants in Sanskrit are left unmarked. It is perhaps for this reason that a scribe decided to introduce a sign for silencing the inherent vowel of a consonant in Sanskrit contexts (see folio below). This sign was included in the proposed block as ◌ZANABAZAR SQUARE SIGN VIRAMA. In the approved encoding for Zanabazar Square, the VIRAMA has the additional function of controlling the production of conjunct stacks.

The VIRAMA was introduced into the proposed repertoire at a late stage during research on the encoding for Zanabazar Square. In fact, the ◌SUBJOINER character proposed here is not altogether a new concept for the script as it was included in earlier tentative repertoires. In the first preliminary proposal for the script, the character that is now called ◌ZANABAZAR SQUARE FINAL CONSONANT MARK was given the name 'NULL VOWEL' on account of its usage for marking a bare consonant (see L2/10-411). The name was later changed to 'VIRAMA' in a revised preliminary proposal (see L2/11-162). As the orthography for conjuncts and the

model for encoding them became better understood, a method was needed for producing conjunct stacks. Although various vowel-silencing marks in Brahmi-based scripts have a secondary function of controlling conjunct formation, it did not seem practical to burden the syllable marker ◌𑨳 in a similar fashion. Instead, a new control character 𑩇 was introduced for this purpose (see L2/11-379). This 𑩇 was given the name 'VIRAMA' and ◌𑨳 was renamed as 'FINAL CONSONANT SIGN'. After discussions with scholars and research on the encoding model for the Soyombo script, the 𑩇 was renamed 'SUBJOINER' and ◌𑨳 was renamed as 'FINAL CONSONANT MARK' (see L2/13-068 and L2/13-198). The discovery of a true ◌𑨴 *virāma* offered the potential to align the script with the Brahmi model. The sign ◌𑨴 was assigned the name 'VIRAMA' and endowed with both vowel-silencing and conjunct-producing properties (see L2/14-024). This approach eliminated the need for the 𑩇 'SUBJOINER'. Of the characters described above, the final, approved repertoire contains ◌𑨳 ZANABAZAR SQUARE FINAL CONSONANT MARK and ◌𑨴 ZANABAZAR SQUARE SIGN VIRAMA.

The removal of 𑩇 'SUBJOINER' and the merger of its function with ◌𑨴 'VIRAMA' may have been a bit too idealistic. It is certainly valid, but it complicates the encoding for two reasons.

First, the dual function of ◌𑨴 ZANABAZAR SQUARE SIGN VIRAMA requires that the conjunct-stacking behavior of the character be suppressed in order to display it visibly when it occurs in a conjunct. Consider a case from the folio below (from Byambaa Ragchaagiin 2005: 113–114).



The VIRAMA occurs twice in the sentence 𑨮𑨂𑨡𑨢𑨴‌𑨰𑨴𑨙𑨂𑨰𑨺𑨭𑨒𑨍𑨙𑨢𑨴 *śubhamstu-sarva-jagatam* "may there be auspiciousness throughout the world" (line 3, column 3). It is used first in the phrase 𑨮𑨂𑨡𑨢𑨴‌𑨰𑨴𑨙𑨂 *śubhamstu*, in the syllable 𑨢𑨴‌𑨰𑨴𑨙𑨂 *mstu*. This syllable has a consonant cluster *-mst-*, of which the 𑨢 MA is marked with a visible VIRAMA and the 𑨰 SA and 𑨙 TA are rendered as a conjunct stack. The second occurrence is at the end of the sentence in the word 𑨒𑨍𑨙𑨢𑨴 *jagatam*, where the VIRAMA is visible and marks the bare consonant 𑨢𑨴 *m*. The usage and shape of VIRAMA provide sufficient proof that it is distinct from ◌𑨳 ZANABAZAR SQUARE FINAL CONSONANT MARK.

To represent the syllable *mstu* as written in the folio using the approved model, it is necessary to block the conjunct-producing behavior of VIRAMA so that the sign is displayed visibly. This is achieved by placing the generic control character ZWNJ U+200C ZERO WIDTH NON-JOINER (abbreviated ZWNJ) after VIRAMA in the encoded sequence:

𑨢𑨴‌𑨰𑨴𑨙𑨂 𑨢 MA, ◌𑨴 VIRAMA, ZWNJ ZWNJ, 𑨰 SA, ◌𑨴 VIRAMA, 𑨙 TA, ◌𑨂 VOWEL SIGN UE

If ZWNJ is not used, then the VIRAMA will produce a conjunct stack:

MA, VIRAMA, SA, VIRAMA, TA, VOWEL SIGN UE

The requirement to use ZWNJ complicates what would otherwise be a simpler encoding model. Using the proposed SUBJOINER and the redefined VIRAMA, the desired representation of *mstu*, as well as other possibilities, would be produced as follows, without the need for invisible control characters:

MA, VIRAMA, SA, SUBJOINER, TA, VOWEL SIGN UE

MA, SUBJOINER, SA, SUBJOINER, TA, VOWEL SIGN UE

MA, VIRAMA, SA, VIRAMA, TA, VOWEL SIGN UE

Secondly, the VIRAMA is not an original element of the script designed by Zanabazar. It does not occur in charts and other descriptions of the script. The sign is a scribal innovation and is likely borrowed from Lantsa or Tibetan, specifically for transliterating these scripts. The usage of VIRAMA in the Zanabazar Square text preserves the graphical representation of the Lantsa and Vartu texts in lines 1 and 2, respectively. The orthographic congruity is carried forward in the Tibetan representation, where *śubhamstu* is transliterated into Tibetan as ཤུབྷམྟུ. Here, the cluster *-mst-* is written by marking ཧ U+0F58 TIBETAN LETTER MA with the ྄ U+0F84 TIBETAN MARK HALANTA and separating it from the conjunct stack <ས U+0F66 TIBETAN LETTER SA, ྟ U+0F9F TIBETAN SUBJOINED LETTER TA>.

Unlike the FINAL CONSONANT MARK, the sign VIRAMA is not commonly used. It is essentially a Sanskrit-specific vowel silencer that complements the FINAL CONSONANT MARK used for Mongolian and Tibetan. Encoding it is necessary for enabling complete representation of Zanabazar Square texts. However, the VIRAMA will be used less for marking bare consonants in Sanskrit and substantially more for producing conjunct stacks, which occur with much greater frequency in the sources. Secondly, as the VIRAMA is not a part of the traditional script it is likely to not be readily recognized by users. Therefore, it is more suitable to restrict usage of VIRAMA to silencing the inherent vowel and to employ the generic character for the common purpose of producing conjunct stacks.

The introduction of the ZANABAZAR SQUARE SUBJOINER improves the encoding model for the script. It serves the specific function of producing a conjunct stack by placing, or rather subjoining, the following letter beneath the previous consonant. It possesses no vowel-silencing function. That purpose should be served by VIRAMA, as well as the existing FINAL CONSONANT MARK, depending upon linguistic context. The VIRAMA would behave like other combining marks in the script and would always be displayed visibly. This redefinition aligns the properties of ZANABAZAR SQUARE SIGN VIRAMA with ZANABAZAR SQUARE FINAL CONSONANT MARK, as well as with TIBETAN MARK HALANTA, which does not have any control properties. This will benefit users of Zanabazar Square, who are accustomed to the Tibetan encoding in Unicode, who would expect the same behavior of the Zanabazar Square sign VIRAMA. Moreover, a dedicated SUBJOINER character aligns with the encoding model for Soyombo, a related script of Mongolia employed concurrently by users of Zanabazar Square, which has also been approved for inclusion in Unicode (see L2/15-004R).

3 Character Data

3.1 Character properties

In the format of `UnicodeData.txt`:

```
11A4x;ZANABAZAR SQUARE SUBJOINER;Mn;9;NSM;;;;;N;;;;;
```

3.2 Linebreaking

In the format of `LineBreak.txt`:

```
11A4x; CM               # ZANABAZAR SQUARE SUBJOINER
```

3.3 Syllabic categories

In the format of `IndicSyllabicCategory.txt`. The ◌𑨴 ZANABAZAR SQUARE SUBJOINER is defined as an '`Invisible_Stacker`':

```
# Indic_Syllabic_Category=Invisible_Stacker
11A4x     ; Invisible_Stacker    # Mn     ZANABAZAR SQUARE SUBJOINER
```

The syllabic category for ◌𑨵 ZANABAZAR SQUARE FINAL CONSONANT MARK and ◌𑨺 ZANABAZAR SQUARE SIGN VIRAMA should be redefined from '`Virama`' to '`Pure_Killer`':

```
# Indic_Syllabic_Category=Pure_Killer
11A33    ; Pure_Killer          # Mn     ZANABAZAR SQUARE FINAL CONSONANT MARK
11A34    ; Pure_Killer          # Mn     ZANABAZAR SQUARE SIGN VIRAMA
```

4 Acknowledgments

This project was made possible in part through a Google Research Award, granted to Deborah Anderson for the Script Encoding Initiative, and a grant from the United States National Endowment for the Humanities (PR-50205-15), which funds the Universal Scripts Project (part of the Script Encoding Initiative at the University of California, Berkeley). Any views, findings, conclusions or recommendations expressed in this publication do not necessarily reflect those of Google or the National Endowment for the Humanities.

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:** *Proposal to encode a SUBJOINER for Zanabazar Square*
2. Requester's name: *Script Encoding Initiative (SEI) / Anshuman Pandey (pandey@umich.edu)*
3. Requester type (Member body/Liaison/Individual contribution): *Liaison contribution*
4. Submission date: *2015-12-30*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *x*
   (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):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *x*
      Name of the existing block: *Zanabazar Square*
2. Number of characters in proposal: *1*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary ; B.1-Specialized (small collection) *x* ; B.2-Specialized (large collection) ;
   C-Major extinct ; D-Attested extinct ; E-Minor extinct ;
   F-Archaic Hieroglyphic or Ideographic ; G-Obscure or questionable usage symbols
4. Is a repertoire including character names provided? *yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *yes*
   b. Are the character shapes attached in a legible form suitable for review? *yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   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 (pandey@umich.edu)
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)?
   (General category properties and line-breaking properties are included)

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? *yes*
   If YES explain *See proposal for history, was proposed with the name (VIRAMA) in L2/11-379*
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? *Agata Bareja-Starzyńska (University of Warsaw, Poland)*
   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? *See L2/15-337*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *rare*
   Reference:
5. Are the proposed characters in current use by the user community? *yes*
   If YES, where? Reference: *By scholars of Mongolian culture, history, and linguistics*
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)? *n/a*
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? *yes*
    If YES, is a rationale for its inclusion provided?
    If YES, reference: *Cf. U+1A60 Tai Tham Sign Sakot*
11. Does the proposal include use of combining characters and/or use of composite sequences? *no*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *yes*
    If YES, describe in detail (include attachment if necessary)
    See proposal
13. Does the proposal contain any Ideographic compatibility characters? *no*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: