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

Doc Type: **Working Group Document**
Title: **Proposal to encode two additional Zanabazar Square letters**
Source: **Andrew West**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2018-06-21**

1.   Introduction

The Zanabazar Square script includes a character U+11A3A ZANABAZAR SQUARE CLUSTER-INITIAL LETTER RA ⊡ which represents a special form of the letter RA ⊓ that corresponds to the Tibetan head letter RA (*ra mgo*) and is used when writing Tibetan conjuncts. In the Tibetan script two characters are encoded to represent RA, U+0F62 TIBETAN LETTER RA for the head letter which takes on a special glyph shape above some letters in Tibetan orthography, and U+0F6A TIBETAN LETTER FIXED-FORM RA which is used in Sanskrit transcription in cases when RA does not change shape above another letter.

This document proposes to encode two additional Zanabazar Square characters corresponding to Tibetan head letters LA (*la mgo*) and SA (*sa mgo*). In the Tibetan script the head letters LA and SA do not have a special form, and so are not encoded separately, but in the Soyombo script the letters RA, LA and SA, as well as SHA, have special forms when used as head letters in Tibetan, and separate characters SOYOMBO CLUSTER-INITIAL LETTER RA, LA, SHA and SA are encoded.

Zanabazar Square texts are not consistent with regard to the representation of the head letters LA and SA in Tibetan syllables, but there is evidence that LA and SA may be written using special vertically-compressed ligatured forms as head letters in Tibetan syllables, in contrast to usage in Sanskrit transcription when these two letters are normally written as standard-form letters, as shown below.

	RA + KA	**LA + KA**	**SA + KA**
Sanskrit Syllable			
Tibetan Syllable			

In standard vertical clusters used for transcribing Sanskrit each consonant letter in the stack is normally full-height or nearly full-height, with a visible gap between the letters. However, in a Tibetan vertical cluster with an RA, LA or SA head letter the RA, LA or SA is vertically-compressed and normally ligates with the base consonant below it, which is also compressed so that both letters together occupy the same vertical height as a single consonant letter by itself. Examples of such vertically-compressed ligatured forms of RA, LA and SA above KA or GA are illustrated in most samplers of the Zanabazar Square script, for example Fig. 1 below (see Figs. 4 through 8 for additional examples).

Fig. 1: L2/15-337 Fig. 1 (C)
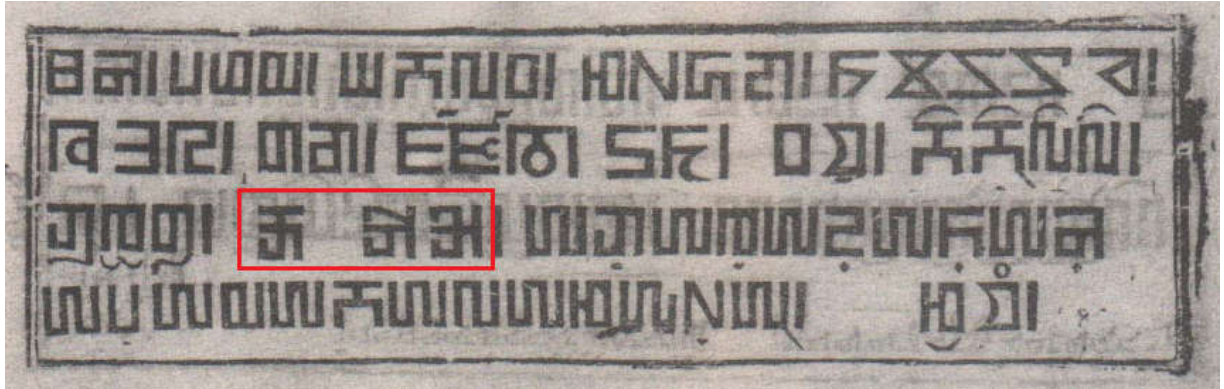
rka, *ska*, and *lka* = Tibetan ཀ་ སྐ་ ལྐ་

Compare the above example with the example below where Sanskrit text is written in Zanabazar Square script. Here the conjunct letters are much higher than single letters, and the *sa* and ta in the highlighted conjunct *staṁ* are both nearly full-height. See Figs. 9 through 11 for other examples with SA + consonant conjuncts in Sanskrit transcription.

Fig. 2: WG2 L2/15-337 Fig. 40 (C)
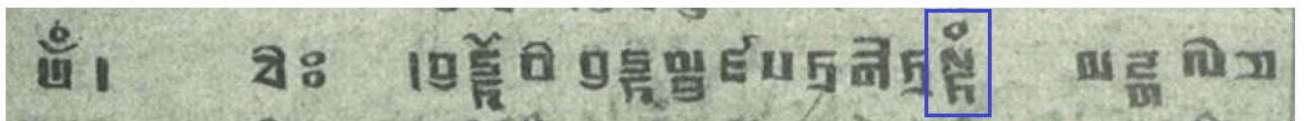
It has to be admitted that the available Zanabazar Square texts are not consistent, and in places show vertically-compressed and ligatured LA and SA head forms in Tibetan syllables, but elsewhere show standard-height, unligatured letterforms (for example Figs. 12 through 15). Actually, usage of cluster-initial RA is also not well-defined, as all the examples of cluster-initial RA given in L2/15-337 show the reduced ligatured form, both in Tibetan conjuncts and in Sanskrit conjuncts (see Fig. 16), and there are no examples of the full-form RA in cluster-initial position.

Anshuman Pandey's proposal for encoding Zanabazar Square script discusses three methods for vertically sizing multiple characters in a vertical cluster (see p. 23-24):

L2/15-337 p. 23
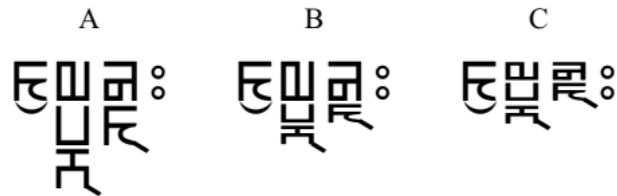
The default method involves no size changes and uses the regular forms of letters (column 'A'). In some sources, the regular size of the initial letter is used, while non-initial letters are compressed along the vertical axis so that their x-height is halved. (see figure 4 and column 'B'). A third practice is to vertically condense the all letter glyphs so that the height of the stack matches the height of surrounding letters (see figure 23 and column 'C'). Depending on x-height such size adjustments may be practical only for stacks of two letters.

It seems to me that Method C is not a general method, but normally applies only to Tibetan head letters RA, LA and SA. In Fig. 23 that he gives as an example of this method, there is a colophon in Tibetan language written in the Zanabazar Square script, and two out of the three clusters highlighted have the vertically-compressed head-letter SA. My analysis is that these two vertically-compressed SA letters represent the special head form of the letter SA, and are not indicative of a general method compressing conjuncts. The first highlighted example is an anomaly, as it shows the syllable *gling* which would normally be written as 𑨀, but is unusually written with a half-sized letter LA instead of U+11A3D ZANABAZAR SQUARE CLUSTER-FINAL LETTER LA.

Fig. 3: WG2 L2/15-337 Fig. 23
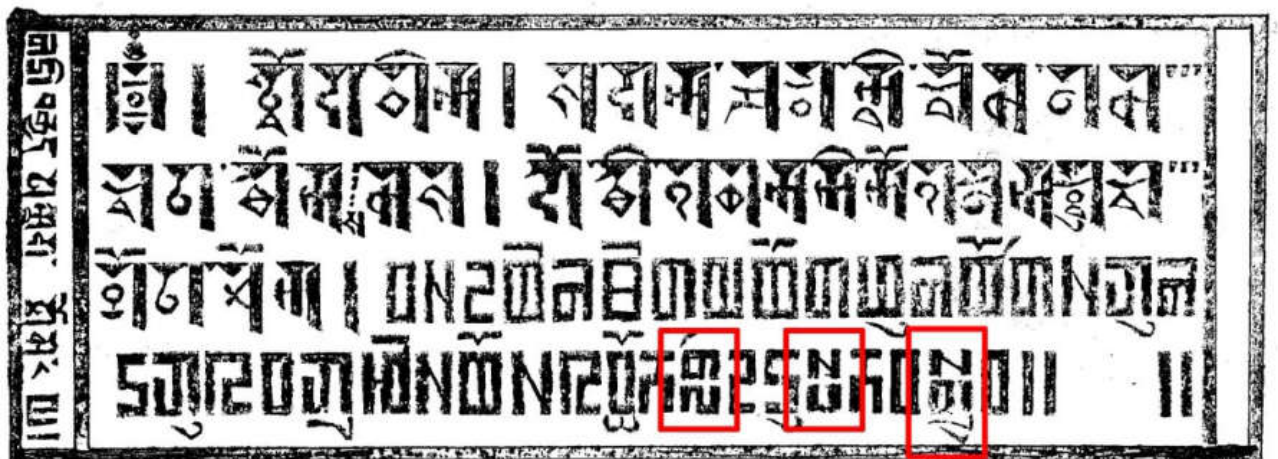
bkra shes chos 'byor *gli*ng *du* *spar* *bsgrub* = Tibetan བཀྲ་ཤེས་ཆོས་འབྱོར་གླིང་དུ་སྤར་བསྒྲུབ

"completed at the Trashi Chöjorling Monastery"

2. *Proposed Additions*

A modern computer font cannot reflect the ambiguity of glyph forms for LA and SA as head letters evidenced in Zanabazar Square texts, and a font designer needs to be able to represent cluster-initial letters consistently. I recently designed and released a Unicode Zanabazar Square font (http://www.babelstone.co.uk/Fonts/Zanabazar.html), perhaps the only fully-functional Unicode Zanabazar font to have been developed to date, and I had to deal with the thorny issue of how to draw the letters LA and SA at the top of a consonant stack. If I followed the Zanabazar Square proposal (L2/15-337) then in all cases I would use the same glyph forms of the letters LA and SA when used as base consonants, with a gap between the LA or SA and the base consonant below, so there would be no distinction between LA/SA plus consonant in a Tibetan stack or in a Sanskrit stack. However, I want to emulate the many examples that show vertically-compressed ligatured forms of LA and SA matching the cluster-initial letter RA in Tibetan stacks. But there is no way of distinguishing Tibetan usage from Sanskrit usage in all cases, so if I use the vertically-compressed ligatured forms of LA and SA at the top of a consonant stack they will be incorrect when transcribing Sanskrit. Currently the only solution is to provide two different versions of the font, one tailored for Tibetan use and one tailored for Sanskrit use, which is not an acceptable solution given that Tibetan and Sanskrit frequently occur in the same text written in Zanabazar Square script. The only reasonable solution is to encode two additional characters: ZANABAZAR SQUARE CLUSTER-INITIAL LETTER LA and ZANABAZAR SQUARE CLUSTER-INITIAL LETTER SA. This would mirror the solution for head letters RA, LA and SA in the Soyombo script. Therefore I am requesting the encoding of the following two characters at U+11A48 and U+11A49 (with U+11A3A shown for comparison):

Code Point	Character Name	Glyph	Corresponding Base Letter	
11A3A	ZANABAZAR SQUARE CLUSTER-INITIAL LETTER RA		U+11A2B	
11A48	ZANABAZAR SQUARE CLUSTER-INITIAL LETTER LA		U+11A2C	
11A49	ZANABAZAR SQUARE CLUSTER-INITIAL LETTER SA		U+11A30	

Unicode data properties:

```
11A3A;ZANABAZAR SQUARE CLUSTER-INITIAL LETTER RA;Lo;0;L;;;;;N;;;;;
11A48;ZANABAZAR SQUARE CLUSTER-INITIAL LETTER LA;Lo;0;L;;;;;N;;;;;
11A49;ZANABAZAR SQUARE CLUSTER-INITIAL LETTER SA;Lo;0;L;;;;;N;;;;;
```

All other properties should be the same as for U+11A3A.

3.   *Recommended Representation of Conjuncts*

As mentioned above, the examples adduced in this proposal are not entirely consistent in the treatment of the head letters RA, LA and SA, which probably reflects the fact that the Zanabazar Square script was never widely used, so a consistent orthographic tradition never developed.

It should also be noted that some of the woodblock prints show better quality calligraphy than others, and the higher quality prints show a more consistent treatment. For example, Figs. 1 and 4 are high quality prints, and both show the compressed and ligated head letters LA and SA. On the other hand, Figs. 3 and 5 are poorer quality, and do not clearly show ligated head letters LA and SA. However, modern scholars who have studied the script all copy the compressed and ligated forms of the head letters RA, LA and SA for writing Tibetan conjuncts, as can be seen in Figs. 6, 7 and 8. In these Tibetan conjuncts the head letters RA, LA and SA are silent, which contrasts with Sanskrit conjuncts where initial RA, LA and SA are pronounced, e.g. *śubhamastu sarvajagatām* in Fig. 10.

In principle, conjuncts for the Sanskrit language should be represented using the normal full-size letters, whereas conjuncts for the Tibetan language should be represented using the cluster-initial letters. However, this is not always the case in practice. We do find examples of Sanskrit conjuncts in Zanabazar script that use the cluster-initial ligatured form of RA, for example *sarva* ᠨᠣ᠊ (= སྥྭ)

and *dharmā* ᠷᠥ᠊ (= དྷྨྲ) in Fig. 16.

For SA and LA, the Sanskrit forms tend to be distinct from the corresponding Tibetan forms:

Sanskrit conjunct *stu*	Tibetan conjunct *sta*
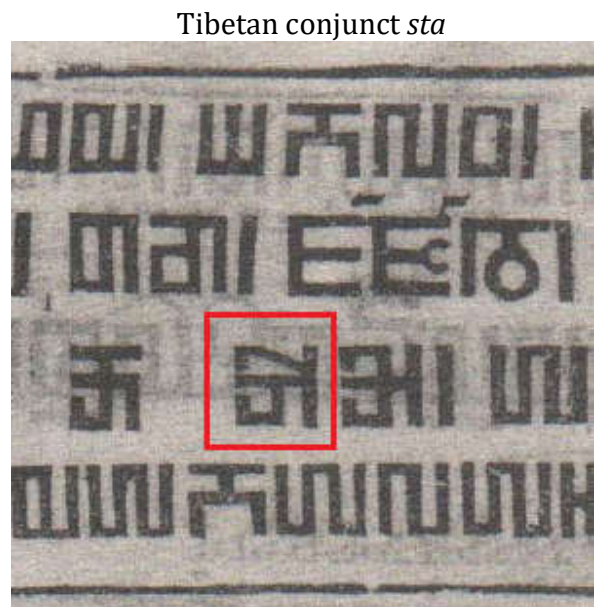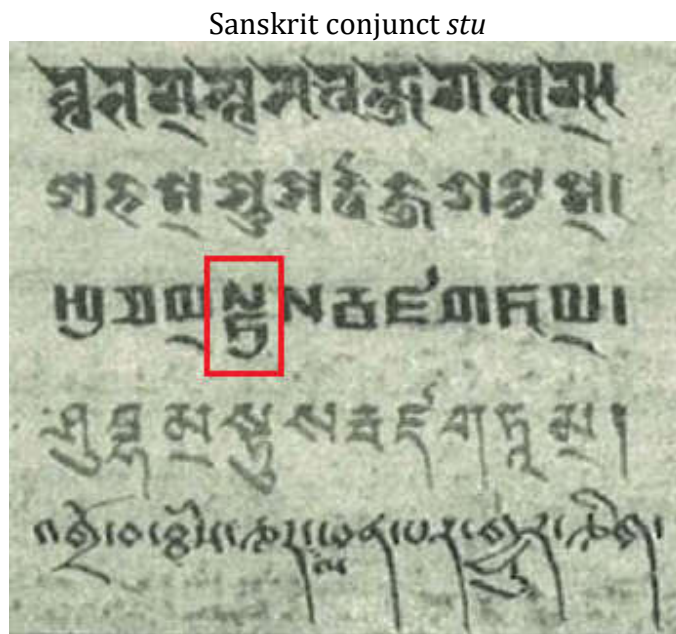
We recommend that Tibetan conjuncts starting with LA or SA should be represented using the proposed cluster-initial letters, whereas Sanskrit conjuncts should be represented using the normal full-size letters. Thus examples such as Tibetan *skyong* ꧀ (=) in Fig. 13, Tibetan *sgrub* ꧀ (=) in Fig. 14, and Tibetan *skad* ꧀ (=) in Fig. 15 should all be represented using the proposed cluster-initial letter SA, whereas examples such as Sanskrit *śubhamastu* ꧀ in Fig.10, and Sanskrit *satva hri disthi tasya* ꧀ in Fig. 9 should be represented using ordinary letter SA + subjoiner + letter TA or THA.
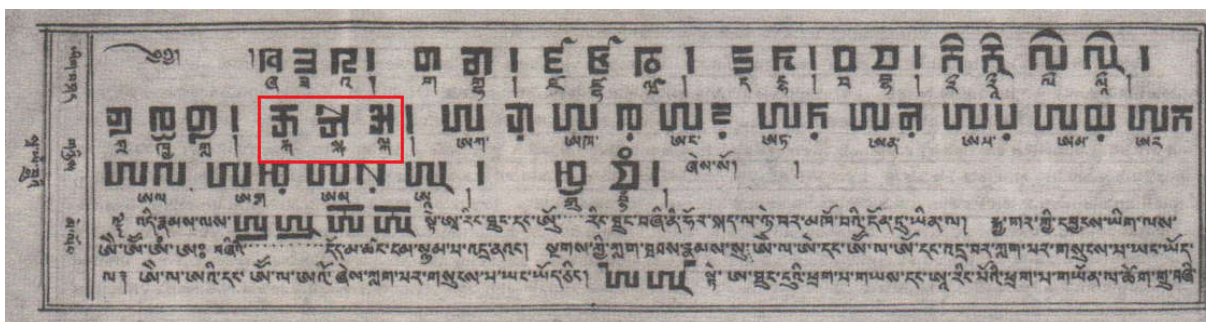
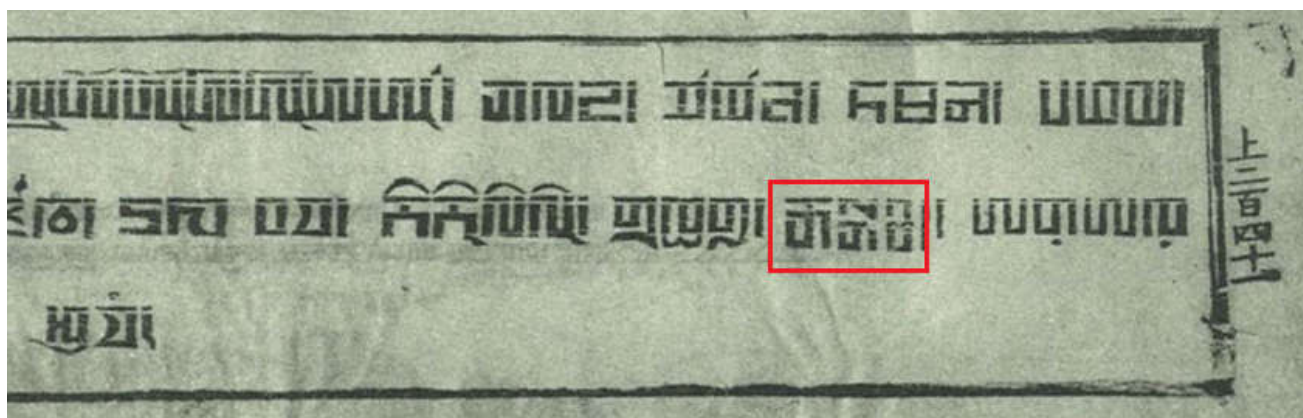## 4. Additional Examples

Fig. 4: WG2 L2/15-337 Fig. 2 (C)

Fig. 5: WG2 L2/15-337 Fig. 5 (B)

Fig. 6: WG2 L2/15-337 Fig. 9

Fig. 7: WG2 L2/15-337 Fig. 10

Fig. 8: WG2 L2/15-337 Fig. 10

52. **tl.** tib., sans. rK(a);
tc. tib., sans. rka. Төвед, самгард хэлний нийлмэл (буюу ra толгойт) rka гийгүүлэгчийн бие даасан буюу (IF) хэлбэр.

53. **tl.** tib., sans. sK(a);
tc. tib., sans. ska. Төвед, самгард хэлний нийлмэл (буюу sa толгойт) ska гийгүүлэгчийн бие даасан буюу (iF) хэлбэр.

54. **tl.** tib., sans. lK(a);
tc. tib., sans. lka. Төвед, самгард хэлний нийлмэл (буюу la толгойт) lka гийгүүлэгчийн бие даасан буюу (IF) хэлбэр.

Fig. 9: WG2 L2/15-337 Fig. 30 (A)

*disth*i (Sanskrit transcription)

Fig. 10: WG2 L2/15-337 Fig. 30 (B)
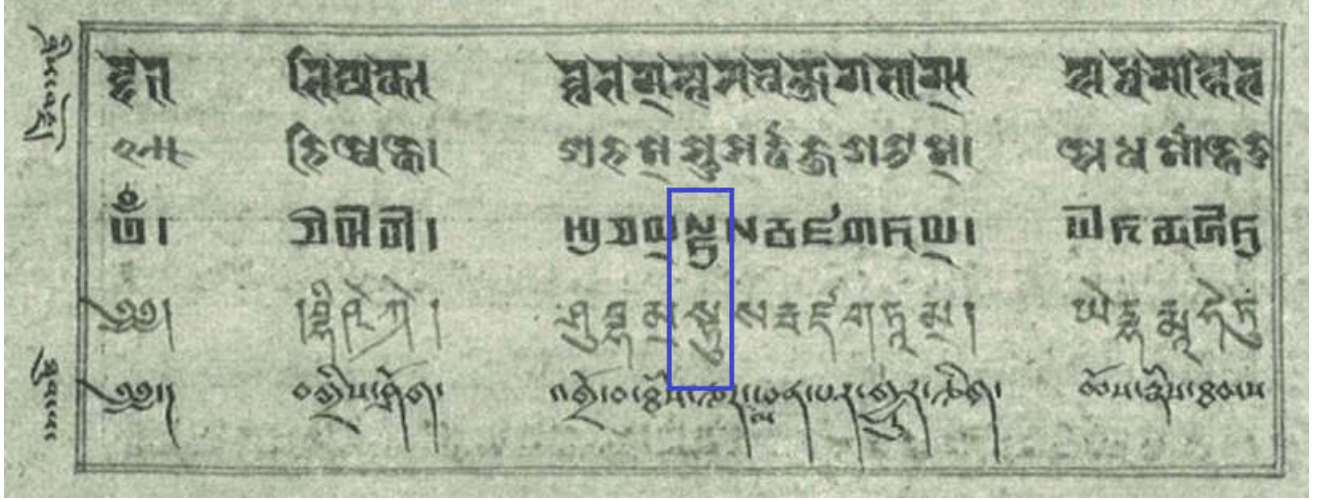
*śubhama**stu** sarvajagatām* (Sanskrit transcription)

Fig. 11: WG2 L2/15-337 Fig. 39
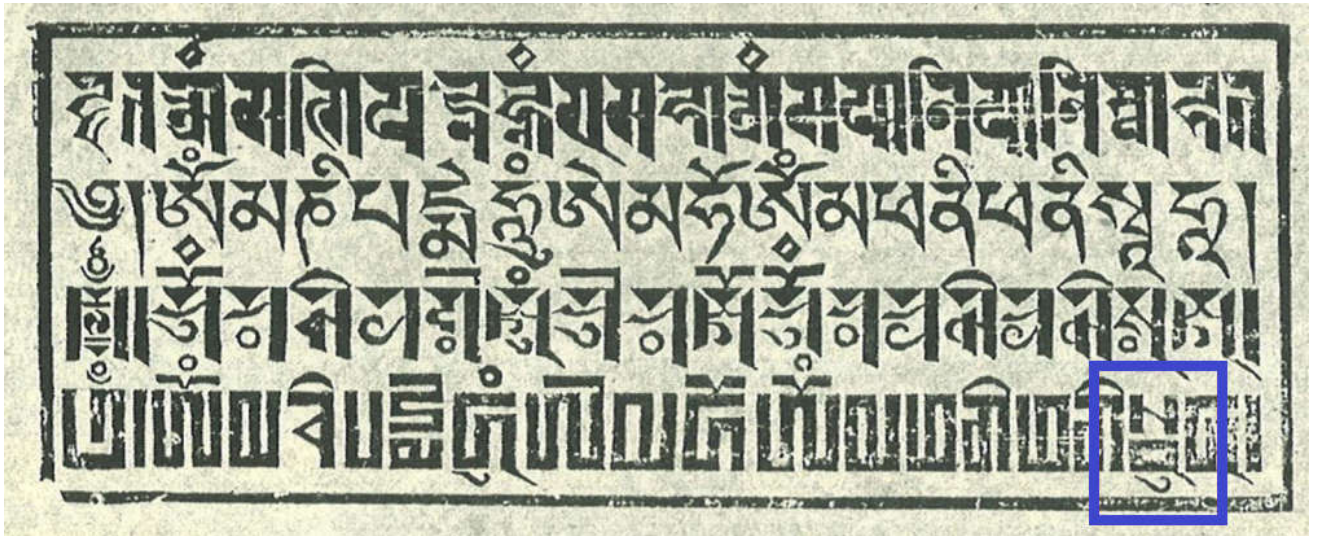
svūhā = svāhā (Sanskrit transcription)

Fig. 12: WG2 L2/15-337 Fig. 3 (D)
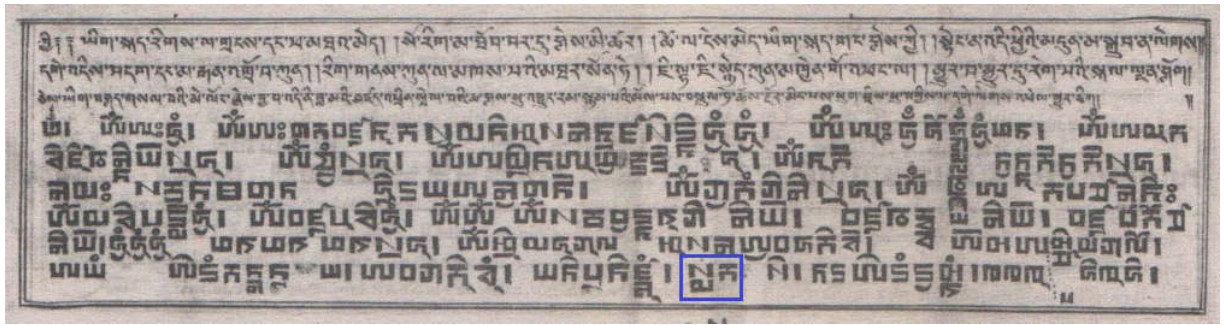
smar = Tibetan སྨར་

Fig. 13: WG2 L2/15-337 Fig. 3 (E)
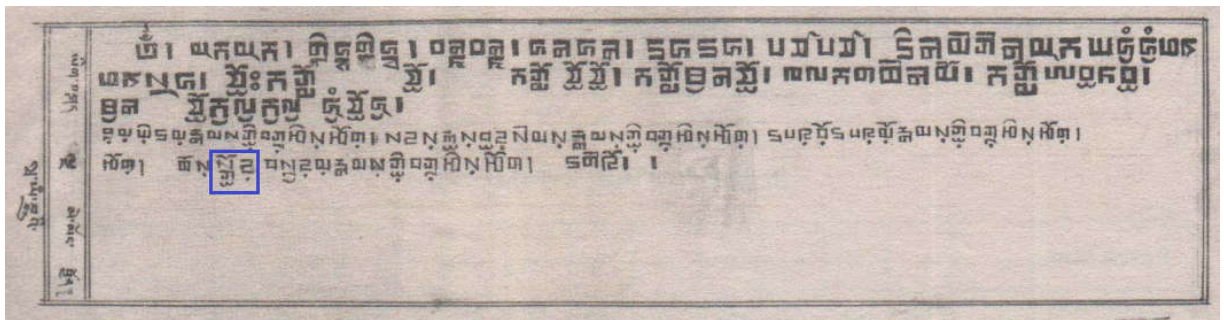
skyong = Tibetan སྐྱོང་ "to protect"

Fig. 14: WG2 L2/15-337 Fig. 40 (D)
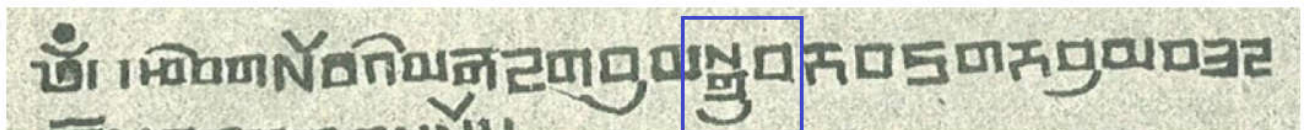
sgrub = Tibetan སྒྲུབ་ "to complete"

Fig. 15: WG2 L2/15-337 Fig. 3 (E)
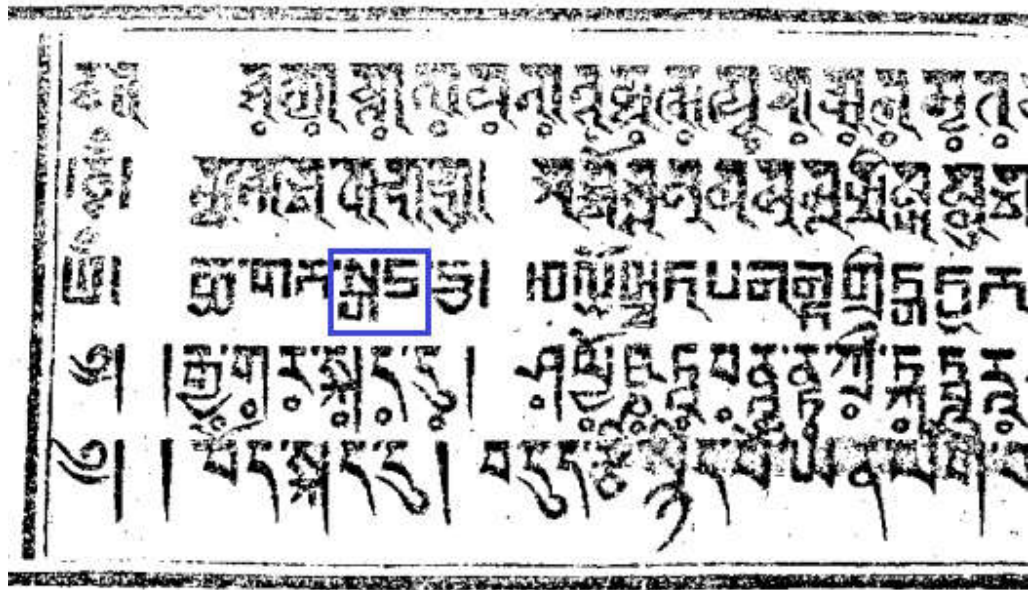
skad = Tibetan སྐད་ "language"

Fig. 16: WG2 L2/15-337 Fig. 30 (B)

sarva ས་ཏ and *dharmā* དྷརྨ

5. *Proposal Summary Form*

<table>
<tr><td colspan="2" align="center">
SO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for guidelines and details before filling this form.
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest Roadmaps.
</td></tr>
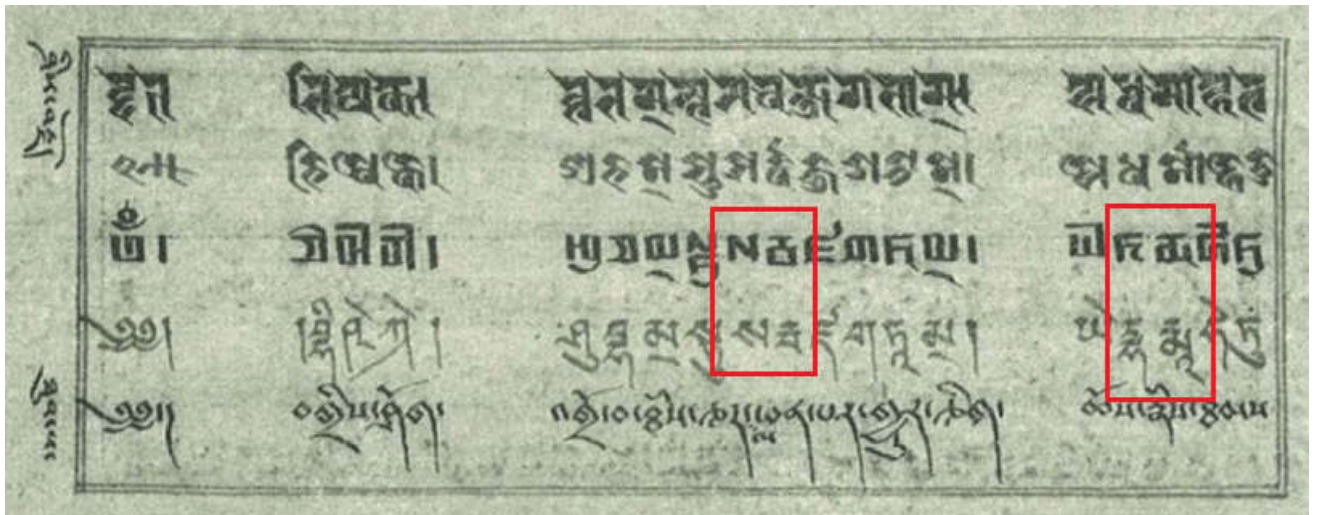</table>

A. Administrative

1. **Title:**	*Proposal to encode two additional Zanabazar Square letters*
2. Requester's name:	*Andrew West*
3. Requester type (Member body/Liaison/Individual contribution):	*Individual contribution*
4. Submission date:	*2018-04-20*
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):	*NO*
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	*YES*
Name of the existing block:	*Zanabazar Square*
2. Number of characters in proposal:	*2*

3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary   **X**   B.1-Specialized (small collection)   B.2-Specialized (large collection)
C-Major extinct   D-Attested extinct   E-Minor extinct
F-Archaic Hieroglyphic or Ideographic   G-Obscure or questionable usage symbols

4. Is a repertoire including character names provided?	*YES*
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?	*YES*
b. Are the character shapes attached in a legible form suitable for review?	*YES*
5. Fonts related:	
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	
Andrew West	
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	
Andrew West	
6. References:	
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	*YES*
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	*YES*
7. Special encoding issues:	
Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?	*NO*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see 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: N4102-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.)?	*NO*
If YES, with whom?	
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?	*NO*
Reference:	
4. The context of use for the proposed characters (type of use; common or rare)	*common*
Reference:	
5. Are the proposed characters in current use by the user community?	*YES*
If YES, where? Reference:	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	*NO*
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	*YES*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	*YES*
If YES, is a rationale for its inclusion provided?	*YES*
If YES, reference:	
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?	*YES*
If YES, reference:	
11. Does the proposal include use of combining characters and/or use of composite sequences?	*NO*
If YES, is a rationale for such use provided?	
If YES, reference:	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	
If YES, reference:	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	*NO*
If YES, describe in detail (include attachment if necessary)	
13. Does the proposal contain any Ideographic compatibility characters?	*NO*
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	