| | |
|---|---|
| TO: | Deborah Anderson, UTC |
| DOC TYPE: | Working Group Document |
| STATUS: | Individual contribution |
| FROM: | Peter Lofting |
| SUBJECT: | **Comments on Tibetan Double Shad** |
| DATE: | 2024–12–27 |

**Summary**

Kushim Jiang proposes clarification of the Core Specification description of shad usage and deprection of U+0F0E ༎ TIBETAN MARK NYIS SHAD.

The present Core Specification description is confusing and needs to be improved; but the double shad should not be deprecated as it has a different meaning.

Arguments and manuscript examples are given to show that double shad behaves as a distinct entity corresponding to the Sanskrit double danda, which is richly represented in Unicode.

A revised Core Specification description of shad usage is proposed.

# ARGUMENTS FOR DOUBLE SHAD

**Chinese data interchange**
Double shad is already encoded in China National Standards <u>GB 29275-2012</u> and <u>GB/T 22238-2008</u>. These are the pre-composed stack encodings for Tibetan which include a mirror copy of the Unicode Tibetan block at the same code points. The U+0F0E code point is therefore needed for lossless round-trip mapping of Tibetan text data between Unicode and these encodings currently used in China.

**Bhutanese data interchange**
The double shad was was first encoded at the request of the Bhutan Dzongkha Development Commission (DDC) in 1988 in the Druk Macintosh. The Unicode Tibetan encoding includes double shad as part of support for this Bhutanese Government code page.

**Double shad is historically and functionally Sanskrit double danda**
The double shad is the Tibetan version of the Sanskrit double danda, which is distinct from the single danda. It represents a longer pause at the end of a text section or verse. The danda is encoded 76 times across 35 Unicode script blocks – see Unicode Tech Note 33 (TN33) LIST OF DANDAS IN THE UNICODE STANDARD https://www.unicode.org/notes/tn33/  UTN33 describes the relationship of Tibetan shad to danda:

> "Many of the danda characters encoded in the Unicode Standard have the word "DANDA" in their name, but there are many instances where punctuation marks are encoded, which historically and functionally are dandas, but which have distinct names specific to a particular script. For example, in Tibetan and scripts influenced by Tibetan, these marks have "SHAD", rather than "DANDA" in their names."

> "…in some cases—such as Tibetan, in which the danda is termed a shad—the danda mark may accrue additional adornments."

The thirty five scripts supporting danda are: Ahom, Balinese, Bhaiksuki, Brahmi, Chakma, Cham, Devanagari, Dives, Javanese, Kaithi, Kawi, Kayah, Kharoshthi, Khmer, Khojki, Lepcha, Mahajani, Meetei, Modi, Mro, Multani, Myanmar, Newa, Ol, Phags-Pa, Philippine, Saurashtra, Sharada, Siddham, Soyombo, Tai, Thai, Tibetan, Tulu-Tigalari and Zanabazar.

Only 7 of these 35 scripts (20%) have a single representation of Danda. The other 28 scripts (80%) encode both single and double danda separately; and four of the 28 scripts have richer representations than that – namely: Tibetan (8 chars), Cham (4 chars), Kawi (3 chars) and Tai Tham (3 chars).

**Sanskrit transliteration data interchange**
Because double shad corresponds to 28 other double danda characters in Unicode, the U+0F0E code point is needed for lossless round-trip mapping of Sanskrit written in Tibetan to these 28 other scripts. This supports the activity of converting the large corpus of Indian sutras that have survived only in Tibetan back into Sanskrit written in these 28 other scripts.

**Existing data**
There are 28 years worth of Tibetan text data accumulated since the U+0F0E code point was designated in Unicode 2.0, released in July 1996. Deprecating U+0F0E would break this legacy data and require that it be processed to decompose the character.

**Existing keyboards, fonts and applications**
Hardware keyboards, fonts and text handling applications would have to be revised for no gain in functionality.

**Mixed semantics**
Kushim Jiang's issues about spacing come from the view that he is looking at the same kind of punctuation strokes and therefore expects to be able to modify their spacing. Once the semantic difference is accepted, the issues of spacing raised become inapplicable, as splitting the double shad would be an over-decomposition that would cause the loss of a distinct semantic particle. Furthermore, splitting the double shad into two single shads creates the risk of the two halves getting separated and thereby losing their meaning.

**Character semantics are independent of display positioning**
The encoded repesentation of single and double shad are based on semantics and are not dependent upon font appearance, as the metrics of any font can be edited to make either character sequence display with the same or different spacing: Some fonts make the width of two single shads equal to one double shad; other fonts distinguish them by making the spacing of two single shads wider than one double shad. The latter is preferable and recommended. See David Corbett's comments on different fonts "Confusion regarding the glyph of U+0F0E TIBETAN MARK NYIS SHAD"
https://www.unicode.org/L2/L2019/19124-pubrev.html

**Manuscripts distinguish double shad from two single shads**
Examination of Tibetan writing from different eras shows that the double shad is always written distinctively close together, in contrast to two single shads, which are always spaced wider apart.

The spacing of two single shads is always at least 2.5 times wider than the spacing of a double shad and often a lot wider than that. The fact that two single shads never get written closer together than an easily recognized margin appears to be in order to avoid confusion between the two kinds of marks. This is consistent with double shad being the double danda. It is also consistent with the observation that double dandas mostly appear in texts of Sanskrit origin, such as verses.

## CORE SPECIFICATION RECOMMENDATIONS

**[1] Distinguish the homoglyph sequence in fonts**
Things that look the same but mean different things are problematic. The most prominent historical case is the LATIN CAPITAL O and the DIGIT ZERO. Font designers already take special care to differentiate this homoglyph pair by slashing or dotting the zero; or making the zero narrower.

The Core Specification should call out the homoglyph problem of the two single shad sequence and recommend that fonts set the spacing of two single shads discernably wider than one double shad.

**[2] Classify double shad as a sanskrit mark**
The Core Specification should classify double shad as Sanskrit punctuation and recommend it be used only with texts of Sanskrit origin.

Core Specification, South and Central Asia–II §13.4 Tibetan pps. 635–636

Proposed insertions in red;
Proposed deletions in ~~strikethrough~~.

**Tibetan Punctuation.** The punctuation apparatus of Tibetan is relatively limited. The principal punctuation characters are the *tsek*; the *shay* (transliterated "shad"), which is a vertical stroke used to mark the end of a section of text; the space used sparingly as a space; and two of several variant forms of the *shay* that are used in specialized situations requiring a *shay*. There are also several other marks and signs but they are sparingly used.

The *shay* at U+0F0D marks the end of a piece of text called "tshig-grub". The mode of marking bears no commonality with English phrases or sentences and should not be described as a delimiter of phrases. In Tibetan grammatical terms, a *shay* is used to mark the end of an expression ("brjod-pa") and a complete expression. Two *shays* separated by one or more spaces are used at the end of whole topics ("don-tshan").

Double *shay* U+0F0E corresponds to the Sanskrit double danda and is normally used in texts of Sanskrit origin. See Unicode Tech Note 33 for the list of 76 danda characters encoded in Unicode across the 35 different scripts that are used to write Sanskrit https://www.unicode.org/notes/tn33/

To prevent visual confusion, it is recommended that font designers set the spacing of two single *shays* discernably wider than one double *shay*. This enables content authors to see the difference and prevent unintended text entry; and is in alignment with observed manuscript practice.

~~Because some writers use the double *shay* with a different spacing than would be obtained by coding two adjacent occurrences of U+0F0D, the double *shay* has been coded at U+0F0E with the intent that it would have a larger spacing between component *shays* than if two *shays* were simply written together. However, most writers do not use an unusual spacing between the double *shay*, so the application should allow the user to write two U+0F0D codes one after the other. Additionally, font designers will have to decide whether to implement these *shays* with a larger than normal gap.~~

**[1] Xylograph showing normal shad usage**



= shad + space + shad
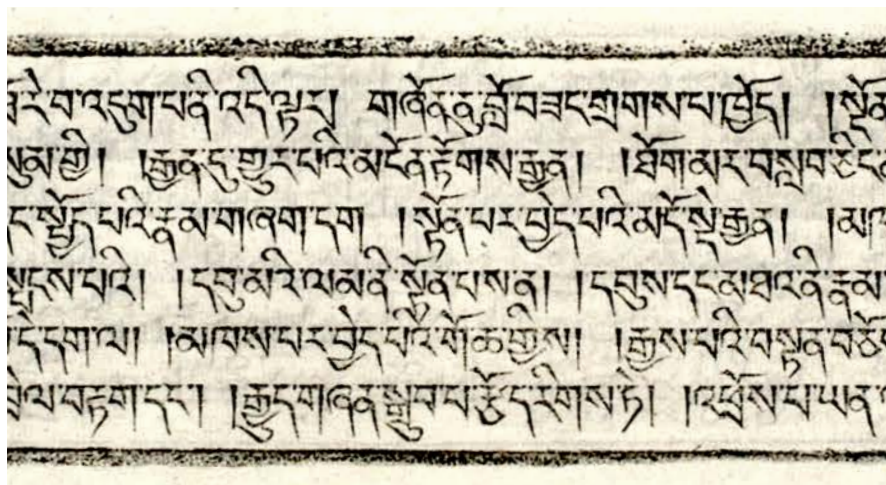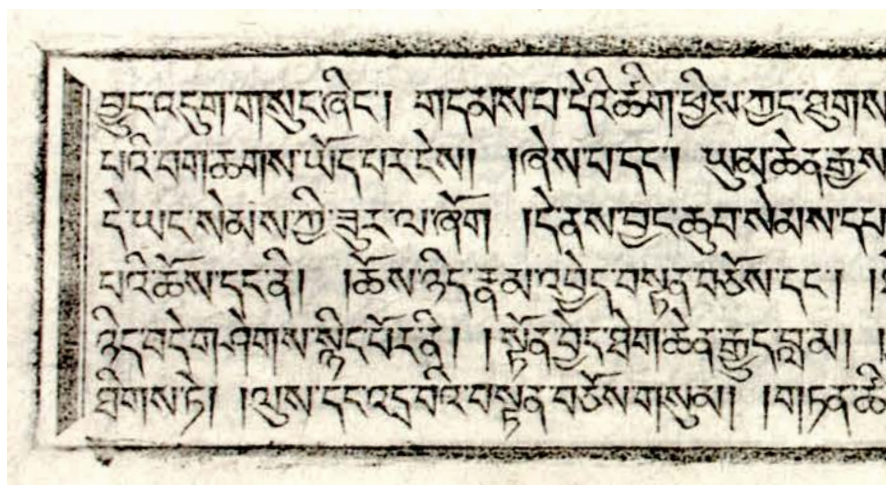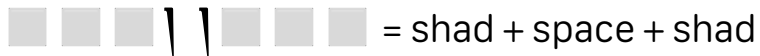




gsung 'bum/ tsong kha pa/ (sku 'bum par ma/)
https://library.bdrc.io/show/bdr:W4CZ302670
Vol.1, img 20

**[2] Xylograph showing double shad usage in a drul shad section break**
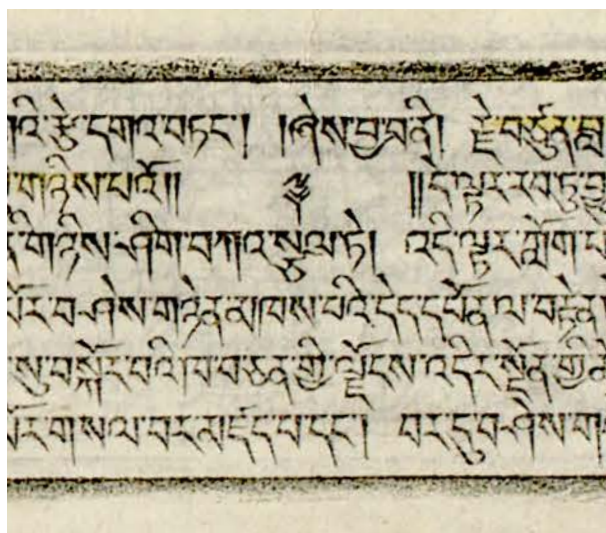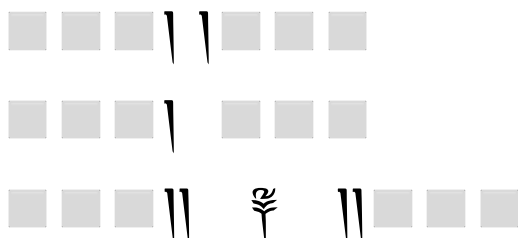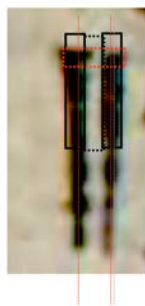






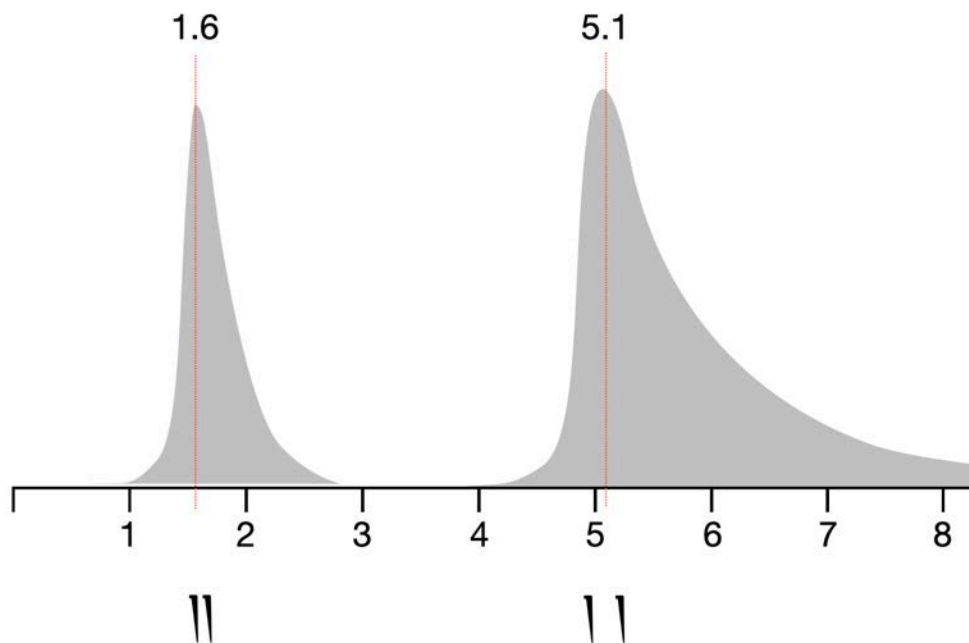gsung 'bum/ tsong kha pa/ (sku 'bum par ma/)
https://library.bdrc.io/show/bdr:W4CZ302670
Vol.1, img 18

**[3] Double Shad & Shad-Space-Shad spacing from the mac OS Kokonor system font**

Kokonor is a facsimile xylograph design.

Units are stroke widths, as measured by the overlay grid.

1.6

5.1
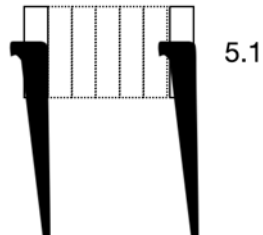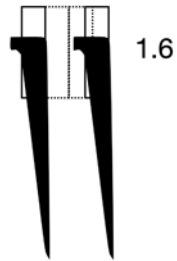
Separation factor = 5.1 / 1.6 = 3.2

1.6

5.1

**[4] Devanagari Double Danda glyphs from freeware fonts**

U+0965 DEVANAGARI DOUBLE DANDA

Akasha

Ancient Sans

Cartoon 1471 Extended

Faulmann Font

Ben Sen Handwriting

Ben Sen

Omniglot Font

Craziest Deco Regular

UnifontEX

The body of the fall (落落の汤圆体)

Lakki Reddy

Charu Chandan Unicode

Charukola Unicode

Modak

Cartoon 1471 Extended

Faulmann Font

Ben Sen Handwriting

Ben Sen

Omniglot Font

OptimaModoki

F77 Minecraft

Chrysanthi Unicode

Code2000

FreeSerif

FreeSans

Annapurna SIL

Pecita

https://www.fontspace.com/unicode/char/0965-devanagari-double-danda

**[5] Devanagari Double Danda – 19th Century hot metal**



spacing = 2.4



तत्राद्यं सर्वमूलं स्यान्मुमुद्दुत्वं न चेतरत् ।
मुमुक्षामन्तरा यत्तु श्रवणं मननादिकम् ॥३५॥
न मुख्यफलंसंयुक्तं केवलं शिल्पवद् भवेत् ।
न शिल्पज्ञानमात्रेण प्राप्यते परमं पदम् ॥३६॥
मुमुक्षामन्तरा यैस्तु श्रुतं सम्यग् विचारितम् ।
शवालङ्कारवत् सर्वं तेषां व्यर्थं भवेत् खलु ॥३७॥
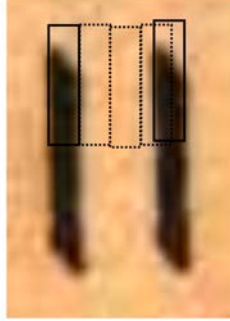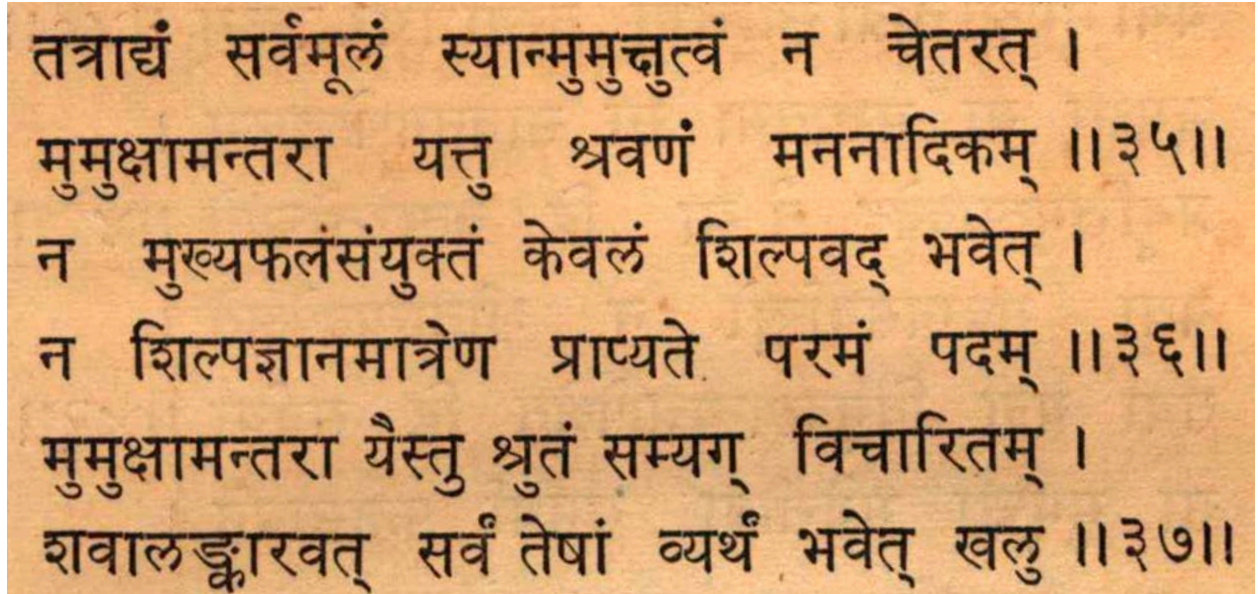
Verses are two lines each:
    First line is punctuated by a single danda.
    Second line is terminated by a double danda.
    Verse numbers are between double dandas

Tripura Rahasyam (Jnana Khanda) Gopinath Kaviraj
https://archive.org/details/TripuraRahasyamJnanaKhandaGopinathKaviraj/page/n13/mode/1up
p.400 of 598

## [6] Xylograph showing mixed single and double shads

༄༅། ། ▮ ▮ ▮

▮ ▮ ▮ ། ▮ ▮ ▮

▮ ▮ ▮ ། ། ▮ ▮ ▮

▮ ▮ ▮ །། ▮ ▮ ▮

▮ ▮ ▮ ། །།
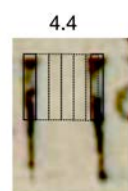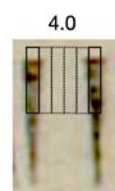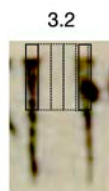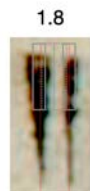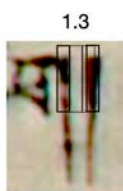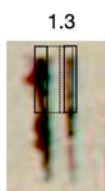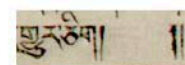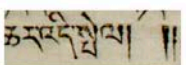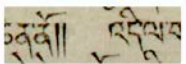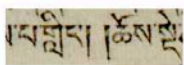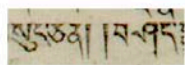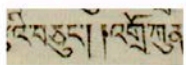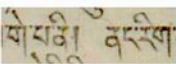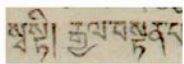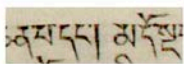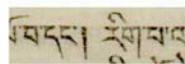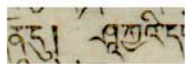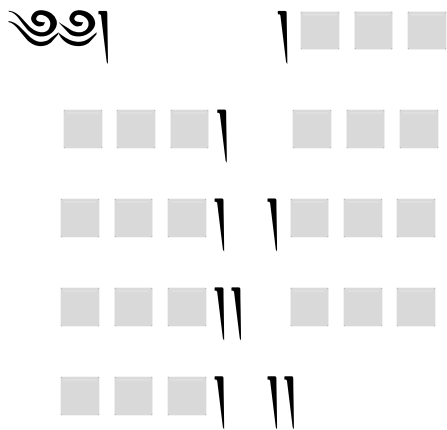


gsung 'bum/ tsong kha pa/ (sku 'bum par ma/)
https://library.bdrc.io/show/bdr:W4CZ302670
Vol.1, img 989

**[7] facsimile xylograph font**



གཅིག་གི་བསྒྱུའི་ཁོངསུ་ལ་གནོད་འཚེ་མི་བྱེད་པར་ཤོག །སྐྱེ་བ་ནས་ཚེ་རབས །

ཐམས་ཅད་དུ་དགེ་བའི་ཚེས་ཀྱི་སྐྱེད་བ་དང་འི་འཕྲལ་བར་ཤོག །

གང་གིས་བདག་ལས་ལེགས་བ་ཤད་ཕོས་པ་འམ།།

མཐོང་ཐོས་དྲན་རེག་གཏམ་གྱི་འཕྲེལ་བཅུན།།

མཚམས་མེད་ལས་བྱང་དན་སོང་སྒོ་ལེགས་ནས།།

ཞིང་མཆོག་པོ་ཏ་ལ་རྒྱེ་བར་ཤོག །

ཅེས་སོགས་གསུངས་སོ།།

སྡད་པ་ལས།

གཙར་ཞིང་གང་ཚག་ལྟ་ཞིང་འགྲོ་ལ་སེམས་འབྱུལ་མེད།།

ཞེས་གསུངས་བ་ལྟར།

དེབ་འདིའི་ནང་གི་ཡིགཆ་དང་རེ་མོའི་སློར་རྒྱལ་སྒྱིའི་ཁྲིམས་ལུགས་ལྟར་སྒྲགས་བཞག་མཛོད་ལུ།

**[8] Manuscript with mixed single and double shads**

༄༅། ° །

། །

། །

།། ।། །།
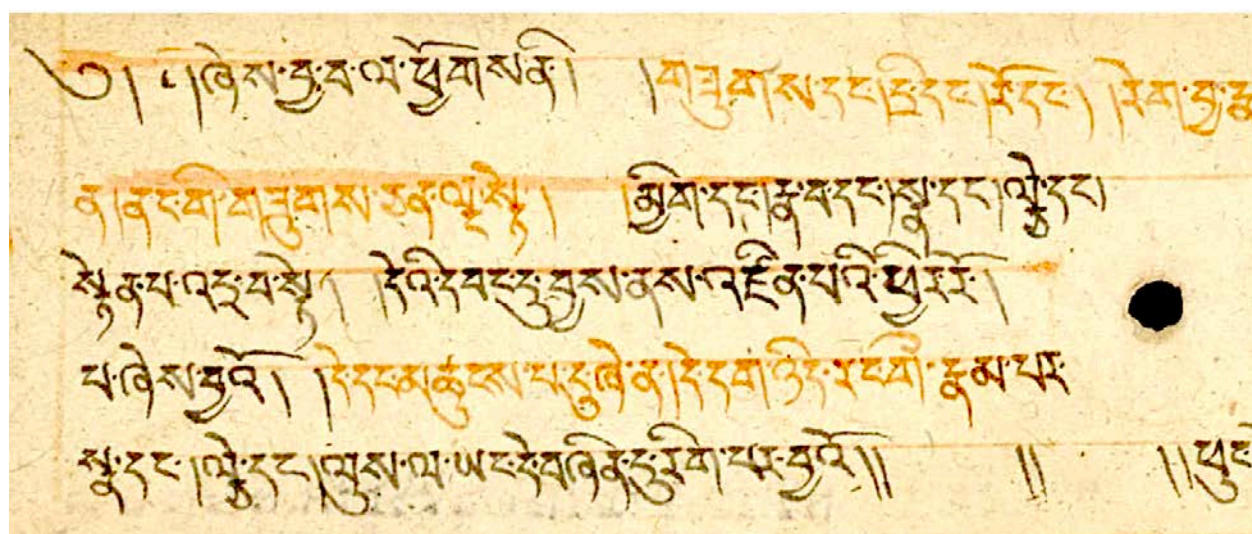
**[8b] Measured spacings of Manuscript with mixed single and double shads**
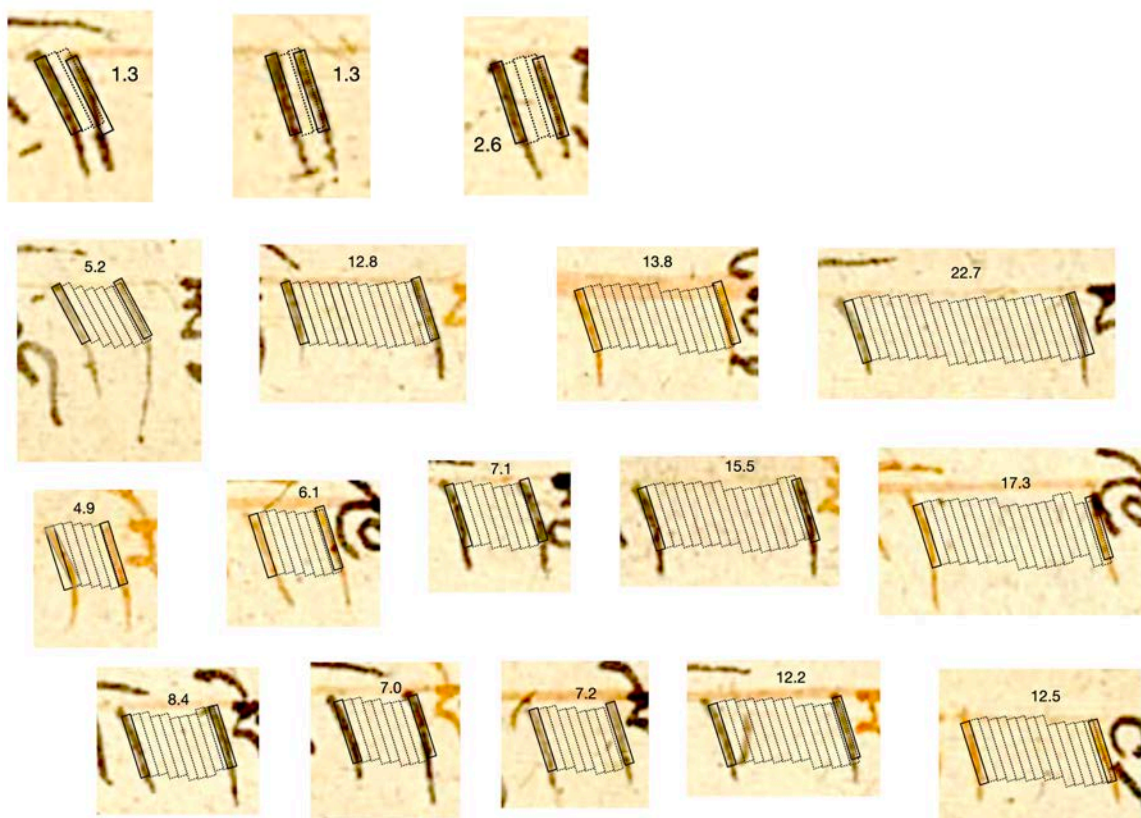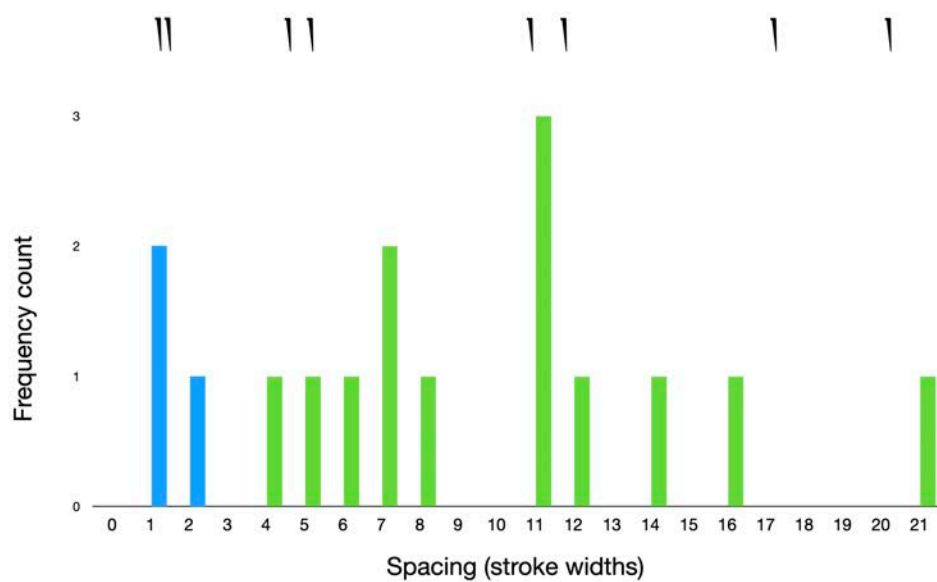
**[9] Manuscript with mixed single and double shads**

ༀ།། ▉ ▉ ▉

▉ ▉ ▉ །  ▉ ▉ ▉

▉ ▉ ▉ ། །  ▉ ▉ ▉

▉ ▉ ▉ །།  ▉ ▉ ▉
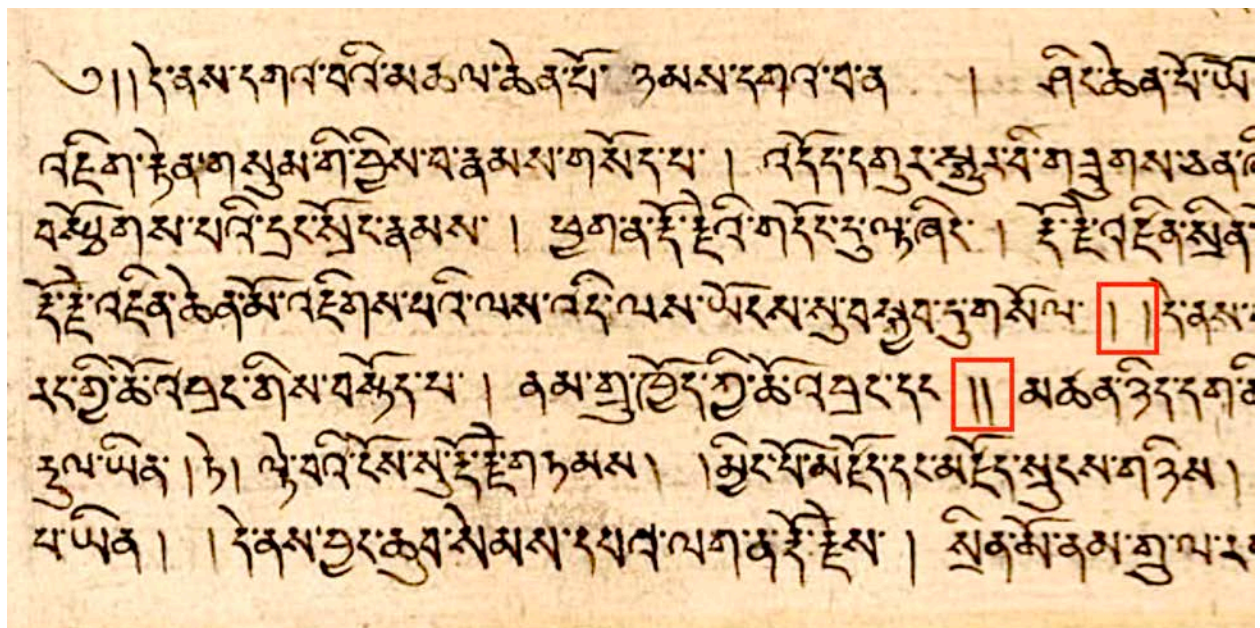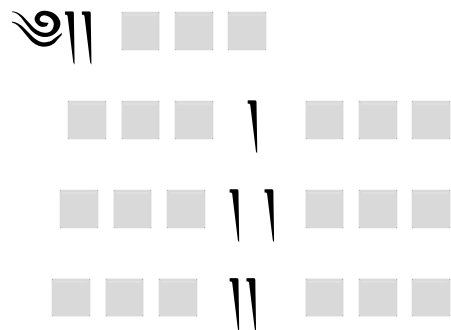


Spacing factor = 6.4/1.6 = 4
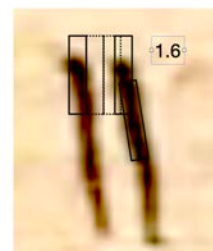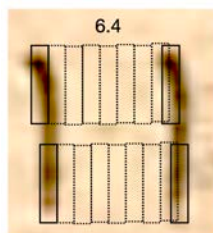
**[10] Manuscript with double shads**

▢▢▢॥▢▢▢॥▢▢▢॥▢▢▢॥

▢▢▢॥ ༀ ॥▢▢▢
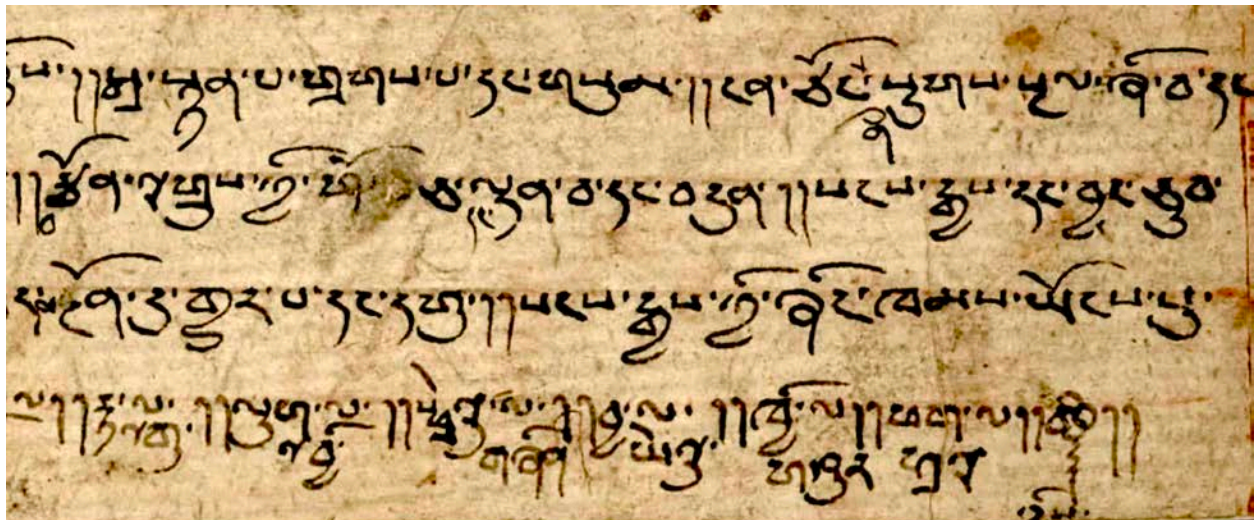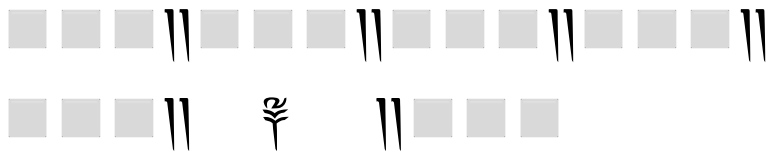
**[10] Manuscript with rubricated shad pairs**

ༀ། ༚ །▢▢▢

▢▢▢▢॥▢▢▢॥▢▢▢॥▢▢▢॥

**[11] Manuscript with mixed single and double shads**

ༀ༎ ༙ ॥ ▨▨▨

▨▨▨▨॥ ॥ ▨▨▨

▨▨▨▨། ། ▨▨▨

▨▨▨▨། ▨▨▨

▨▨▨▨॥ ॥ ॥ ▨▨▨

▨▨▨▨།
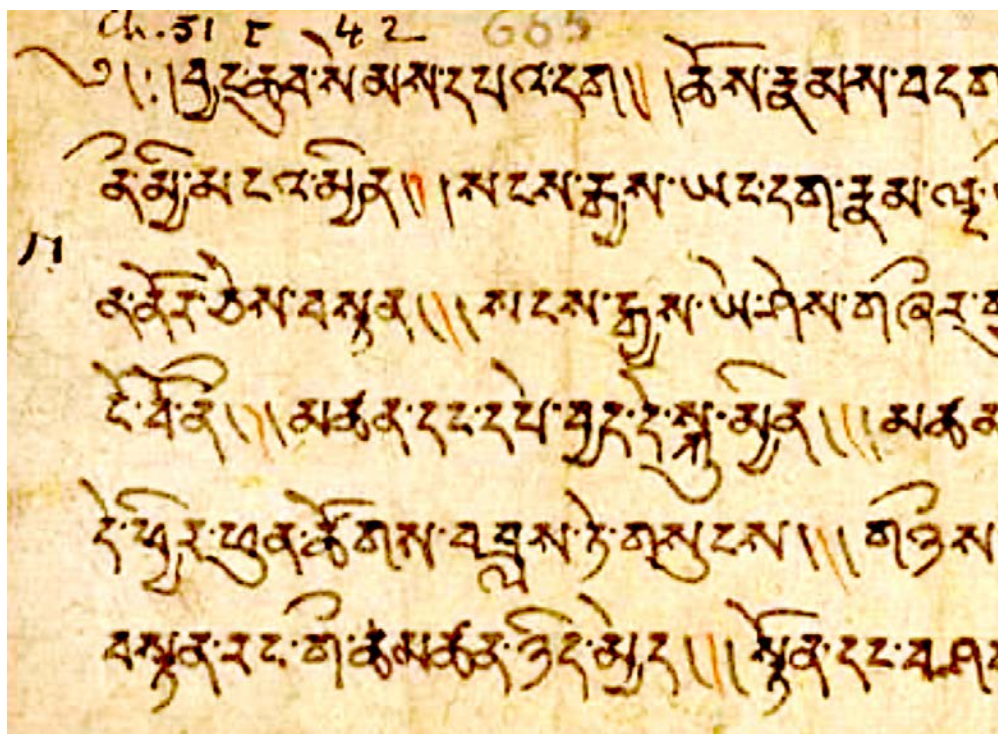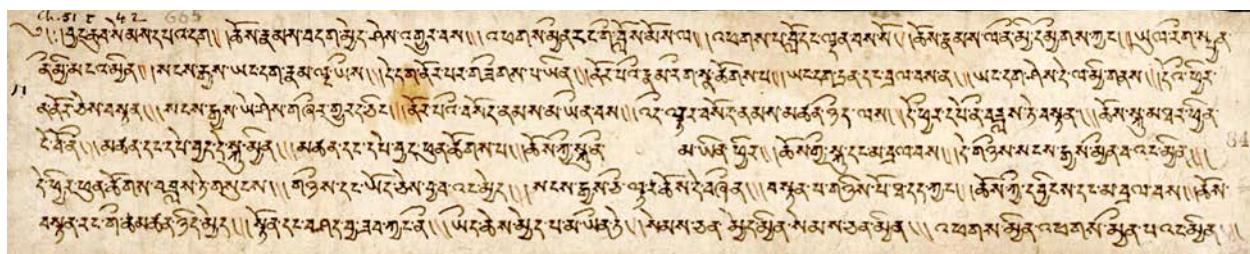


https://idp.bl.uk/collection/7A3CF0D8D3DD49FCA00FA5BCA0ADBE17/?return=/collection/?page=3&term=Tibetan+MS
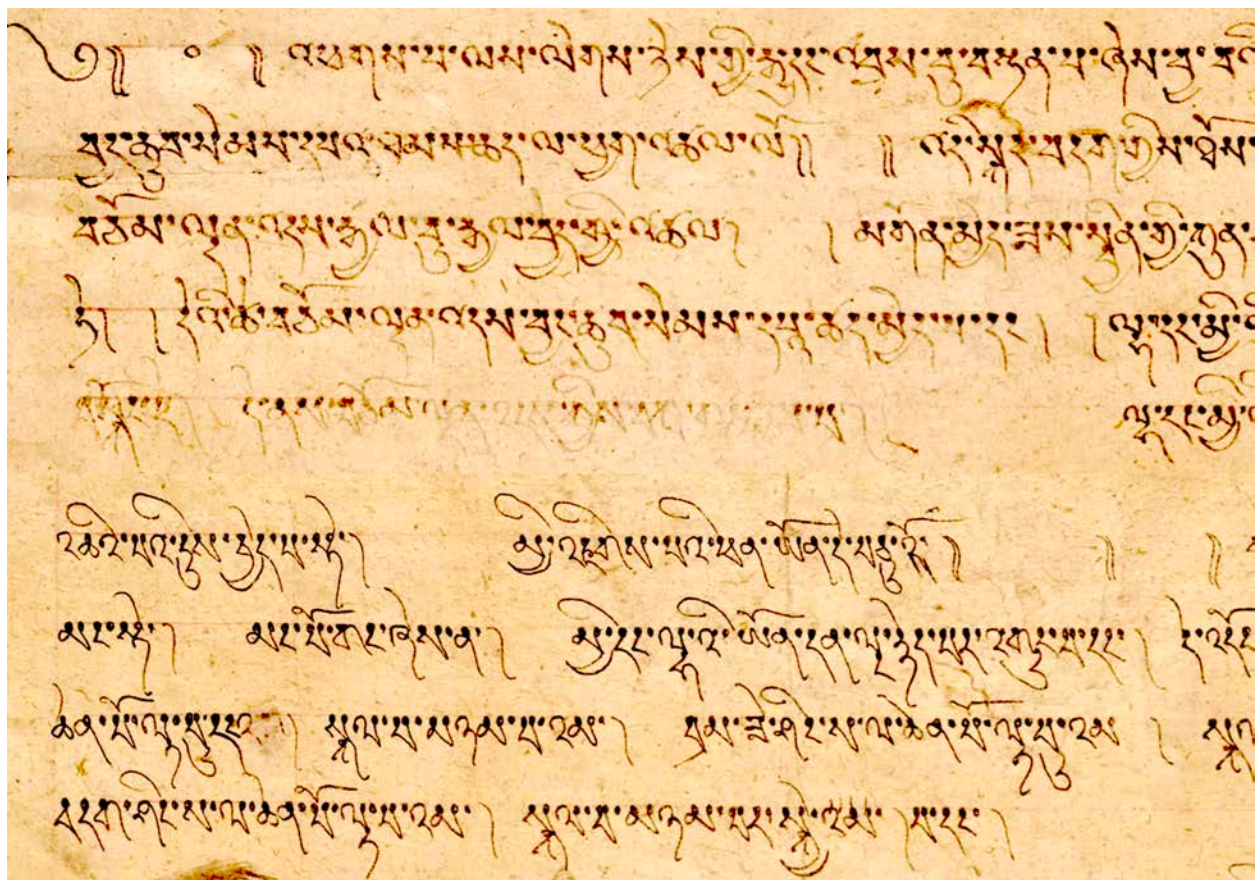
**[12] Manuscript with mixed single and double shads**
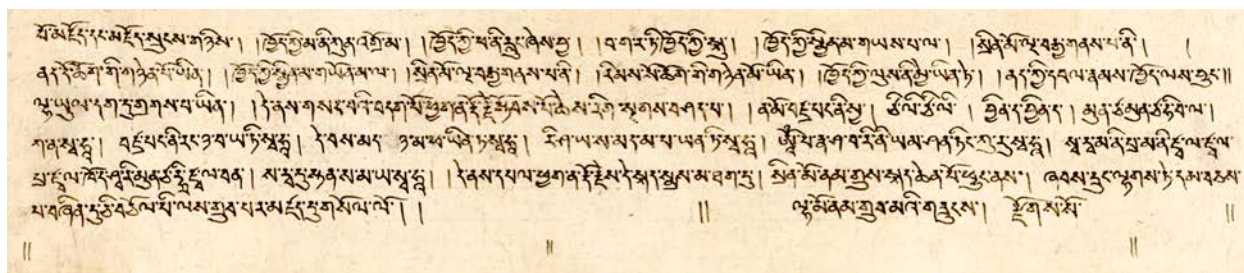
**[13a] Combined punctuation patterns from all samples**

**[13b] Combined punctuation patterns from all samples**

▪▪▪ ❜ ❜ ▪▪▪

▪▪▪ ❜ ❜ ▪▪▪

▪▪▪ ❜　❜ ▪▪▪

▪▪▪ ❜ ❞

▪▪▪ ❞ ❞ ▪▪▪

▪▪▪ ❞ ⚘ ❞ ▪▪▪

▪▪▪ ❞ ❞ ❞ ▪▪▪

▪▪▪ ❞ ❞ ❞ ▪▪▪ ❜ ❜

▪▪▪ ❜ ❜ ❞ ▪▪▪ ❞

❞ ❞ ❞

## [13c] Combined punctuation patterns from all samples

ༀ།། ▪▪▪

ༀ། ༔ །▪▪▪

ༀ། ྅ །▪▪▪

ༀ།། ྅ །། ▪▪▪

ༀༀ། །▪▪▪

▪▪▪ཀ།།

▪▪▪ག།།

▪▪▪ །

▪▪▪།།