L2/25-209

To: Script Encoding Wording Group
From: Peter Lofting
Date: 3 January 2025
Subject: Response from Peter Lofting to Kushim

1 [Text from Peter L]

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.

[comment from Kushim]

I discussed this issue with Academician Nyima Trashi in Beijing, and he argues that the double shay has stable and inherited semantics, so it should not be deprecated.

[response from Peter L]

✅

2 [Text from Peter L]

The danda is encoded 76 times across 35 Unicode script blocks - see UTN #33 which describes the relationship of Tibetan shad to danda.

[comment from Kushim]

It would be a good idea to introduce UTN #33 in the relevant paragraph of the core spec.

[response from Peter L]

✅

3 [Text from Peter L]

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.

[comment from Kushim]

Disagree. The distinct semantic particle has long been lost.

[response from Peter L]

A fair point for many historical Tibetan language texts and modern colloquial Tibetan usage.

But how to reconcile this observation of lost meaning of double shad with Nyima Trashi's and my point that "double shay has stable and inherited semantics"?
It would be up to subject matter experts to identify when text is of Sanskrit origin and should be rendered with double-shad.

Giving a more explicit usage recommendation to avoid double shad might therefore be one way.

From p.4 of my comments – the proposed core spec edits:

> **Two *shays* separated by one or more spaces are used at the end of whole topics ("don-tshan").**

Change this to read:

> **Two shays are normally used at the end of whole topics ("don-tshan"). In Tibetan language texts, shad punctuation patterns should normally be composed of sequences of single shad characters – including double shad marks, which should be typed as two consecutive single shads. Wide pairs of shads and shads with other symbols between them should be typed as single shad followed by one or more spaces or non-breaking spaces and any symbols, then another single shad.**

This would discourage the double shad character from use in colloquial texts, which clears the way for 'line adjustment' to be able to operate on any single shad pattern, including the tightly spaced double-shad marks for those that wish it. But is that the right thing to do?

On the other hand, the double shad character could serve a valuable function in colloquial Tibetan text for freezing the spacing of these closely spaced double marks, while permitting justification code to adjust other pairs of shads. See more below. This could be compared to the handling of double quote marks in Latin. These are logically decomposable into two single quotes, but there is no advantage and it looks cruder. Nearly all word processors today automatically convert two single quotes into single double quote characters.

4. [Comment from Kushim]

> There is a lot of material (including the material provided in the comments) to support that the width in the middle of a double shay is the result of line adjustment

[response from Peter L]
Yes in part; but not all widths in the middle of all double shads, as line adjustment is not all that is going on:

The distribution of widths is not continuous, but clustered: There is a cluster of two closely spaced shads; and then an empty spacing range - a "zone of exclusion" of several stroke widths (2-4+). Then there is a cluster of a "spaced pair" in the 4 to 8 stroke width range which appears to be the most common spacing difference in the samples I looked at. I illustrated it with a hypothetical probability distribution graph on p. 7 and real measured distribution on p. 13 of the comments.

This spacing difference is clear and contrastive. That many different authors contrast the spacing this way must reflect something. What that something is may vary across authors, traditions, languages and eras. But it is not random justification noise in the line layout.

Beyond the common spacing of about 4 stroke widths for the "spaced pair" of single shads, some manuscripts show additional wider spacings, such as the MS on p.13. For anyone interested in examining a larger sample, this image was the recto side of image group 25 of https://idp.bl.uk/collection/679E2BCC8873438DA3F96E83756323F8/?return=/collection/?page=8&term=Tibetan+MS (sorry I left out the page ref in the comments).

It would take lexical content analysis of texts to corroborate if and where these spacing differences were meaningful marks of changes in the lexical structure (verse, section, topic, chapter, great section ends, etc).

5. [comment from Kushim]

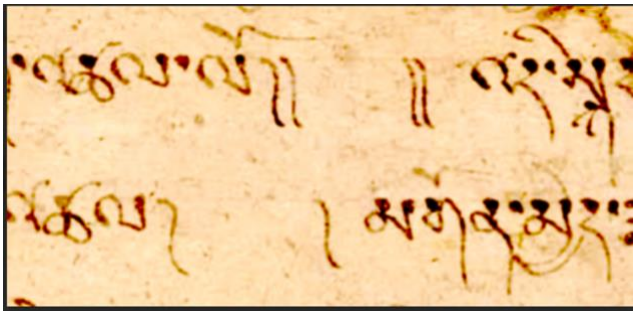
> the width in the middle of a double shay is the result of line adjustment and therefore should not be fixed by type design (comment page 13, page 14),

[response from Peter L]

The two references to my p.13 and p.14 are examples two hand-written manuscripts where I've measured the differences in shad spacing.

I agree that some part of the variation of spaced shad pairs is as Kushim says "…. *that the width in the middle of a double shay is the result of line adjustment.*"

However, I believe the tightly spaced double shad is an exception, as there is only very small variation observed in the tightly spaced shad pairs – the double shads. For the strongest constrasting case look at sample [11] on p.17, where you can see consistently tight double shads in the same line with widely varying spaced shad patterns. Double shads are written consistently tightly. They behave as a different thing to widely spaced shad pairs. They behave like well-written double quote marks.



Kushim shows contemporary Tibetan typeset books in his paper as illustrations of the presences of line layout justification spacing adjustments. He uses this as an argument to propose splitting the double shad marks to enable spacing adjustment between the two strokes. Can Kushim show any manuscript examples where the double shads vary the same amount as spaced pairs of shads?

I'd like to better understand what Kushim wants to achieve and how he sees the tightly spaced double shad versus the spaced shad pair. Does he see these as identical? Are the tight pairs available for line justification in his view? The observed "zone of exclusion" in the spacing distributions in the samples says they are not available for spacing out in this way. I expect that a larger sample of texts would reinforce this pattern.

If Kushim needs single shads everywhere for controlling his publication layouts, he could achieve this goal today by applying an in-house editorial rule to only use single shad characters; and for any text data received from out-of-house, he could do a search and replace to convert all double shads to a pair of single shads.

Other publishers may want a different house style: Those publishers who want to prevent the tightly spaced double shad pair from being widened, while at the same time adjusting all the wide spaced shad pairs, they will need some method of distinguishing the double shads. A special justification algorithm could be used; control character(s) could be embedded; or… a separate encoding could be used.

What other text manipulations does Kushim need to support that require him to treat text in his way?
Does Kushim have a sorting algorithm or search function that needs this uniformity?

6. [comment from Kushim]

and that other symbols can be inserted in the middle of a double shay, which cannot be supported by existing character features (comment page 16).

[response from Peter L]
This is referring to the rubricated spaced shad pairs on p.19.

How to represent this is a great and open question.

Is it a triple shad or a spaced pair of shads decorated with an inserted red mark? Or a display variant of a double shad character with an inserted mark (but no color)?

In terms of original manuscript order of creation, I would guess that the author wrote the text in black ink first and then he or someone else went back over it with a red pen to highlight these parts. The red strokes are not as consistent as the black strokes, which suggests it might have been done in a hurry or by a different author. Perhaps an owner of the text later on highlighting a section for chanting.

What does the rubrication signify? Are they verses? Are they a root text followed by a commentary? Is this a translation of a sanskrit text and these are honorific double dandas? Translation is needed to answer this.

In terms of digital text, it is the most straightforward to represent this as three shad characters with the middle shad having a color style attribute. This is convenient graphically and for typeseting, but probably wrong semantically; because if the post-authoring rubrication theory is subscribed to, these are decorated spaced shad pairs. This becomes a case where semantics and graphics diverge. A compromise to maintain consistent data representation of shad pairs everywhere might be to represent the text electronically by a spaced shad pair which is colored red. Another treatment might be to color the whole text string red. If search and sorting are not a priority, then these would be unnecessary compromises over the original appearance.

With the present state of evolution of Tibetan text data, representing appearance over semantics is preferable in cases like this, because it provides a "Witness" representation that records the appearance of the text. This can be processed and transformed by a series of rules into a semantically consistent "Judgement" representation for indexing and search purposes. The reverse transformation is not possible. So I would argue that the Witness representation takes first priority and the Judgement version for database use be derived from it.

The way Kushim is using this example suggests he is thinking of this BLACK-RED-BLACK triplet as a double shad with a red decoration that is not a shad inserted between the two strokes. Normal fonts can't support color variation within a single glyph. Color can only be applied to complete glyphs as a style attribute.

Would Kushim object to typing this example as three single shad characters?

Most fonts space two consecutive shads the same width as the double shad character, so this supported by the present generation of fonts. It is also how the graphic above was typed.

If a consistent representation of the text semantics is a priority, then having all pairs of shads encoded the same way is a desirable convenience. If enforced, then the above graphic case becomes very hard to represent except by use of custom font variant glyphs or color font formats. That takes it out of the realm of general digital text representation into specialist publishing formats, which makes the text data much less available and versatile.

Given the rich variety of shad punctuation patterns in existence in the corpus, it would seem that a more flexible string indexing and comparing algorithm would take the weight off this issue and enable the diverse range of manuscripts to be accurately represented in a witness form without loss of text processing convenience.

7. [Comment from Kushim]

Comments have already mentioned (page 21) that the semantics of double shay will be provided by single shay after the Tibetan letters ka and ga.

[response from Peter L]

This functionality is supported in fonts today. Contextual substitution rules within a modern OpenType or AAT font can change the double shay when it follows any specified character (the previous character defines the "context" for the shaping rule to activate). In this case the change would be to substitute the double stroke glyph for a single stroke glyph variant. This is represented in pseudocode below by the ".singlestroke" dot suffix on the glyph names.

ka + doubleshad --> ka + doublshad.singlestroke
ga + doubleshad --> ga + doublshad.singlestroke

8 [Comment from Kushim]

And the existing double shay character cannot provide the semantics.

[response from Peter L]

Fonts without shaping rules cannot provide this; but shaping rules can be added to any font today, as described above. Tibetan unicode fonts used today already depend on the presence of hundreds of such shaping rules for normal text display; so adding a few more is not a new class of feature, but is simply a few additions to the large number of existing rules.

This is not the only context-dependent display behavior in Tibetan. For publication quality typesetting, all context-dependent behaviors need to be handled – either by

hard formatting common today, or by automatic contextual substitution which is yet to be worked through comprehensively.

It would be a very worthwhile collaboration to assemble the list of typesetting requirements needed in books published in Tibetan today.

Identifying all these behaviors would enable a more informed discussion of the different ways to control layout and what is most beneficial to include in plain text versus higher level annotations or display algorithms. In other words where the most useful dividing line should be drawb between text data and display.

A starting list is collated by Richard Ishida in the W3C Tibetan Orthography Notes https://r12a.github.io/scripts/tibt/bo#phrase

A lot of the points in the W3C notes are from Tony Duff's Word Tibetan! 5.1 manual which Richard Ishida cites eleven times in his notes. https://collab.its.virginia.edu/access/content/group/26a34146-33a6-48ce-001e-f16ce7908a6a/Tibetan%20fonts/Tibetan%20Legacy%20Fonts/Tibetan!.pdf

From the semantic data processing field, an example use case similar to controlling layout is in Tibetan Natural Langauge Processing, where the need to not break up pairs of shads and to keep space following a single shad is being represented by insertion of U+005F _ LOW LINE. See the sample text segmenting output below, which is generated from the Tibetan word segmenter at https://github.com/OpenPecha/pybo

Text input

"ཞྭ། །རྒྱ་གར་སྐད་དུ། བོ་ཌྷི་སཏྟ་ཨ་བ་ཏཱ་ར། བོད་སྐད་དུ། བྱང་ཆུབ་སེམས་དཔའི་སྤྱོད་པ་ལ་འཇུག་པ། །

སངས་རྒྱས་དང་བྱང་ཆུབ་སེམས་དཔའ་ཐམས་ཅད་ལ་ཕྱག་འཚལ་ལོ། །བདེ་གཤེགས་ཆོས་ཀྱི་སྐུ་མངའ་སྲས་བཅས་དང་། །ཕྱག་འོས་ཀུན་ལའང་གུས་པར་ཕྱག་འཚལ་ཏེ། །བདེ་གཤེགས་སྲས་ཀྱི་སྤྱོམ་ལ་འཇུག་པ་ནི། །ལུང་བཞིན་མདོར་བསྡུས་ནས་ནི་བརྗོད་པར་བྱ། །"

Segmented text output... (2s.)

ཞྭ།_། རྒྱ་གར་ སྐད་ དུ །_ བོ་ ཌྷི་ སཏྟ་ ཨ་བ་ ཏ་ ར །_ བོད་སྐད་ དུ །_ བྱང་ཆུབ་ སེམས་དཔའ་ ཀི་ སྤྱོད་པ་ ལ་ འཇུག་པ །_། སངས་རྒྱས་ དང་ བྱང་ཆུབ་ སེམས་དཔའ་ ཐམས་ཅད་ ལ་ ཕྱག་ འཚལ་ ལོ །_། བདེ་གཤེགས་ ཆོས་ ཀྱི་ སྐུ་ མངའ་ སྲས་ བཅས་ དང་ །_། ཕྱག་འོས་ ཀུན་ ལ འང་ གུས་པ་ ར་ ཕྱག་ འཚལ་ ཏེ །_། བདེ་གཤེགས་ སྲས་ ཀྱི་ སྤྱོམ་ ལ་ འཇུག་པ་ ནི །_། ལུང་ བཞིན་ མདོར་བསྡུས་ ནས་ ནི་ བརྗོད་པ ར་ བྱ །_།

9 [Text from Peter L]

> 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.

[Comment from Kushim]

Disagree. The spacing of two single shays should not be determined by the type design phase, but by text adjusting in typesetting phase.

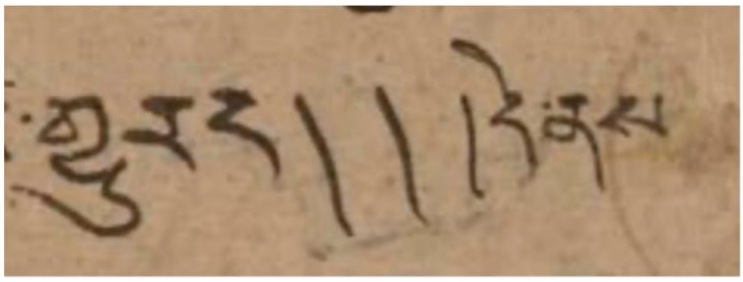
[response from Peter L]
Agreed the separation distance of widely spaced shad pairs should not be frozen and their variation is driven in part by line justification – either manual in manuscripts, or automatic in typeset documents.

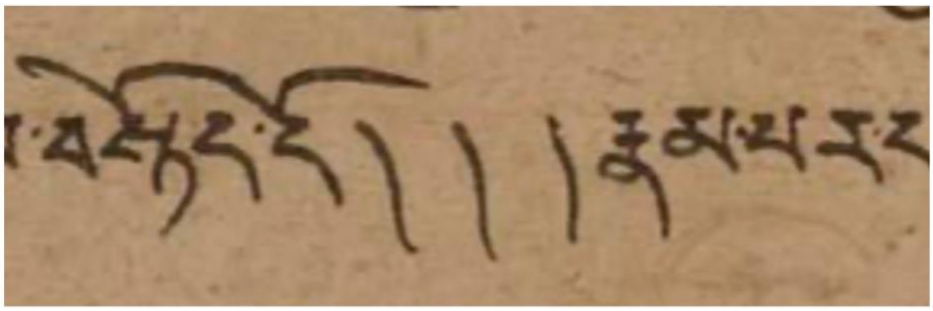
But the separation distance of the closely spaced double shad marks does not change in sampled manuscripts and therefore should not be adjusted by line justification.

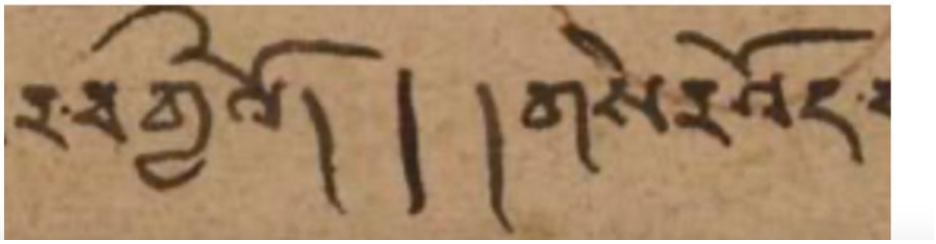
There are also cases of three shads in manuscripts. If three shads are accepted as a punctuation pattern to be supported, then what is their semantic representation? Are they a different class of entity? The natural witness representation would be to type single shads separated by one or more spaces or non-breaking spaces.

Again, this case points at the need for specification of a more flexible Tibetan string sorting and comparison algorithm.
Collecting edge use cases will support development of this.

Here are triple shads in a MS
https://idp.bl.uk/collection/D4777FE7B3764470BADF2D3FEE982070/?return=%2Fcollection%2F%3Fpage%3D3%26term%3DTibetan%2BMS







And another MS with asymmetrical spacing — a pair then a long space then a starting shad
https://idp.bl.uk/collection/3003E7CDA6874C19A109E8F5AC244EFF/?return=%2Fcollection%2F%3Fpage%3D4%26term%3DTibetan%2BMS