L2/25-207

Universal Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по Стандартизации

Doc Type:	Working Group Document
Title:	Proposal to clarify the usage of *shay* in the *Core Specification*
Source:	Kushim JIANG (姜兆勤)
Status:	Individual Contribution
Action:	For consideration by UTC
Date:	2024–10–25

0 Background

Regarding the usage of U+0F0E ། TIBETAN MARK NYIS SHAD, the current *Core Specification* states that:

> ... Two *shays* are used at the end of whole topics (དོན་ཚན་ *don-tshan*). Because some writers use the double *shay* with a different spacing than would be obtained by coding two adjacent occurrences of U+0F0D ། TIBETAN MARK SHAD, 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.

In the Comments on Public Review Issues (L2/19–124), David Corbett reported that:

> ... I've downloaded a bunch of Tibetan fonts and most of them display U+0F0E ། as slightly narrower than two U+0F0D ། Many make them the same width. A few of the Qomolangma fonts make U+0F0E ། slightly wider. The code chart glyph for U+0F0E ། consists of two shays so close together there is barely any space between them. If the standard is correct, the code chart glyph is misleading, if not wrong, and should have more space between the shays. If the majority of my test's fonts are correct, Chapter 13 should not imply their spacing is wrong.

Liang Hai raised the issue with Tibetan Layout Task Force, since the then current version (W3C First Public Working Draft 16 June 2020) of *Requirements for Tibetan Text Layout and Typography* (TLReq) mentions the related issues and recommended ways to handle them. Liang says:

> Currently my personal impression is: This character was encoded as a magical character to allow that magic of spacing out two *shays* to happen at where this character is used. But it's become clear that this character is not very helpful for that matter, because it relies on specialized typesetting environments, and also, the first *shay* for such spacing out situations is often absorbed by the preceding letter's vertical stroke (and this character with two *shays* probably cannot be used).

The relevant content has been deleted and it began to refer to *Tibetan Orthography Notes* after W3C Group Draft Note 05 July 2024, saying:

> ... Topics (e.g. headlines, verses, and longer paragraphs) are often terminated with a double *shad* or separated with *shad* + space + *shad* ... A phrase that ends with the root consonant U+0F40 ཀ TIBETAN LETTER KA or U+0F42 ག TIBETAN LETTER GA will normally swallow up the *shad* that immediately follows it, even if there is a vowel sign. For example, where you might expect to see a double *shad*, you might see ཀུ ། and སྐུ ། However, the *shad* is not omitted if these characters have a subscript, e.g. ལྐ། །

> ... When a phrase ends with *shad* + space + *shad* the space between the *shad* marks is normally reduced in Tibetan pechas, down to 1/4 or 1/3 of the normal width, or made to fit the space available. Some space is retained

to avoid the appearance of a double-*shad*. Boundaries between chapters or significant sections may also be represented by a double-*shad* followed by 5 to 6 spaces and another double-*shad*.

... U+0F0E ༎ TIBETAN MARK NYIS SHAD can be used for the double-*shad*.

... A line that ends with a *shad* plus space followed by a consonant can wrap after the *shad* and discard the space. But a line that ends with one of the following must not lose the space and must not be broken either side of the space: (1) U+0F40 ཀ TIBETAN LETTER KA or U+0F42 ག TIBETAN LETTER GA followed by a space (in which case a *shad* is not used); (2) followed by a space (in which case a *shad* is not used). This should be straightforward if content authors use U+00A0 NO-BREAK SPACE for the latter cases.

With this background, this proposal clarifies the behavior of *shad* (U+0F0D) and *nyis shad* (U+0F0E) and requests that the relevant paragraphs of the *Core Specification* be updated.

1 Introduction

In Tibetan text, the graphemes that may be analyzed as being represented by *shads* can be divided into two categories, one located at the beginning of a sentence and one at the end of a sentence. Roughly speaking, at the beginning of a sentence, *shad* may need to be used with, for example, U+0F04 ༄ TIBETAN MARK INITIAL YIG MGO MDUN MA, U+0F05 ༅ TIBETAN MARK CLOSING YIG MGO SGAB MA; at the end of a sentence, *shad* may appear on its own or be used with other decorative characters, for example, U+0F08 ༈ TIBETAN MARK SBRUL SHAD.

Starting Indicator. The literature of the Tibetan Empire period is dominated by the Dunhuang literature, which contains the literature of Bön and Buddhism. Bön literature usually starts with ༄༎ or ༄། །; Buddhism literature usually starts with ༄༎ or ༄། །, but also starts with ༄།༔།. By the Era of Fragmentation, ༄༅། and ༄༅ྃ། began to appear in abundance, followed by ༄༅༎, ༄༅༅༎, ༄༅༅༅༎ and ༄༅༅༅༎ in the 11th and 12th centuries. By the 13th century, the starting indicators were essentially unified, including ༄།, ༄༅།, ༄༅༎, ༄༅༅༎, ༄༅༅༅༎, ༄༅༅༅༎, ༄༅༅༅༅༎, ༄༅ྃ།, ༄༅ྃྃ།, and ༄༅ྃྃྃ།. And there were also ༄ཱཱུྃ།, ༄ཱཱུྃ། and ༄ཱ།.

In modern Tibetan literature, both ༄༅༎ and ༄༅། ། are used and only either ༄༅༎ or ༄༅། ། appears in the same book.

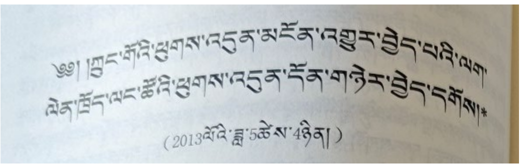

Fig. 1 Page 68 of [Xi, 2015a], showing ༄༅། །

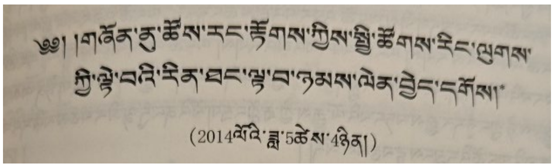

Fig. 2 Page 231 of [Xi, 2015b], showing ༄༅། །

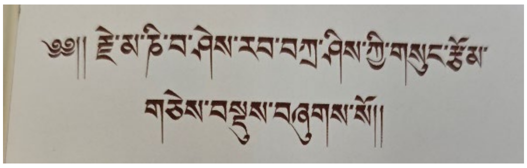

Fig. 3 Cover page of [Suiduo, 1999], showing ༄༅༎

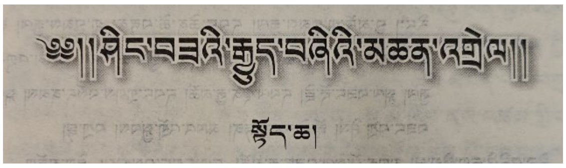

Fig. 4 Cover page of [Shingbzav, 2018], showing ༄༅༎

Ending Indicator. The ending indicators in Tibetan literature include *chig shad* (ཚེག་ཤད, །), *nyis shad* (ཉིས་ཤད, ༎), *bzhi shad* (བཞི་ཤད, ༎ ༎), *sbrul shad* (སྦྲུལ་ཤད, ༈) and so on. Some ending indicators also appear as a combination of *shads* and decorative symbols, such as with the *rgyan shad* (རྒྱན་ཤད, ༑) to form ༎༑༎, ༎༑༑༎, or with the *sbrul shad* (༈) and the *rnam bcad* (ཿ) to form །༈།, །ཿཿ།, ༎ཿཿ༎, ༎ཿ༑ཿ༎ and so on.

In the vast majority of the modern Tibetan literature, only the *chis shad* (ཚེག་ཤད, །) will be used for the ending indicator. In some of the more formal modern Tibetan literature, both ༎ and ། ། are used and only either ༎ or ། ། appears in the same book.

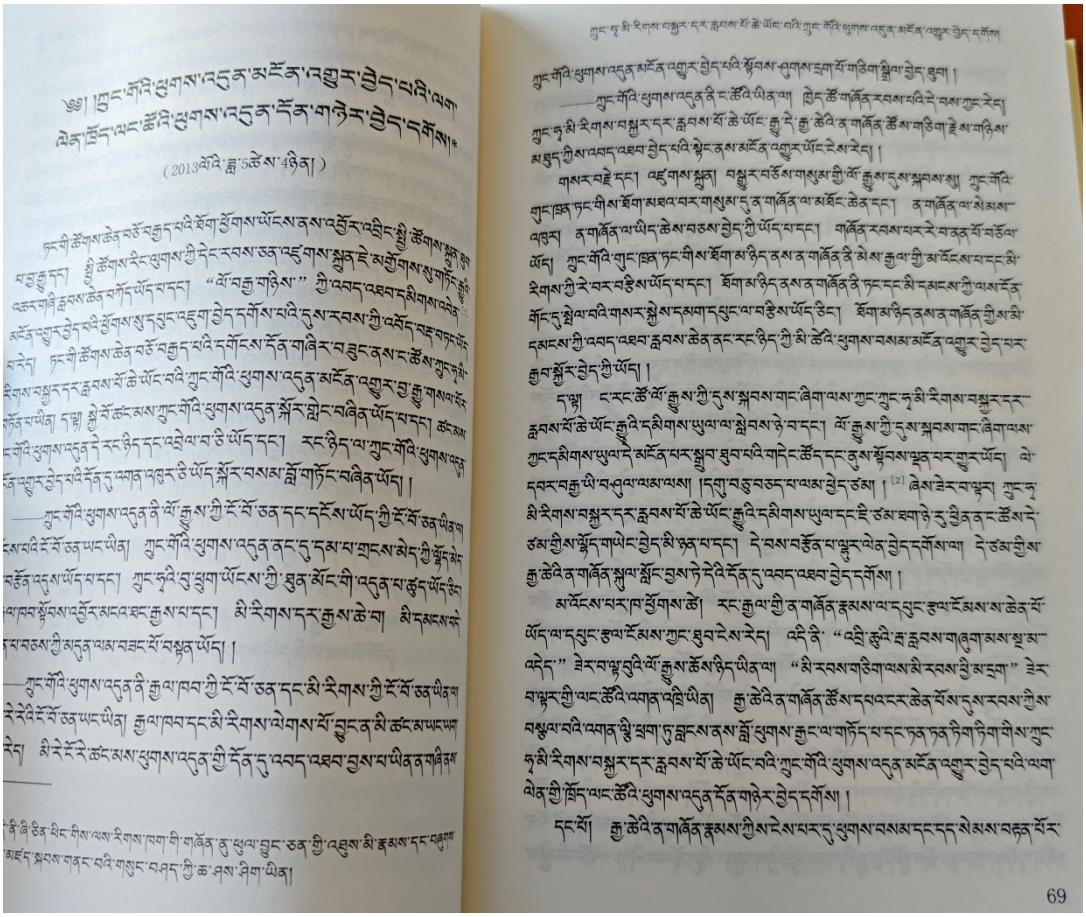

Fig. 5 Page 68 and 69 of [Xi, 2015a], showing | |

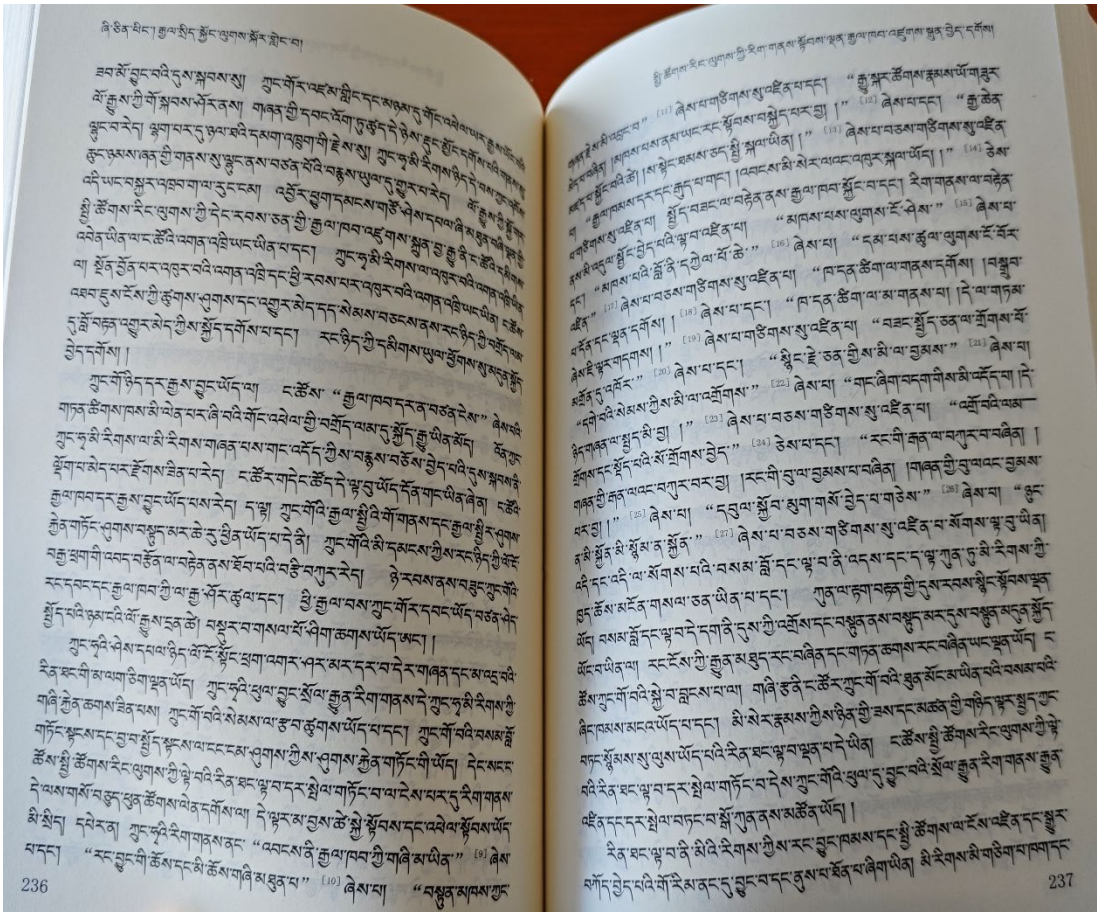

Fig. 6 Page 236 and 237 of [Xi, 2015b], showing | |

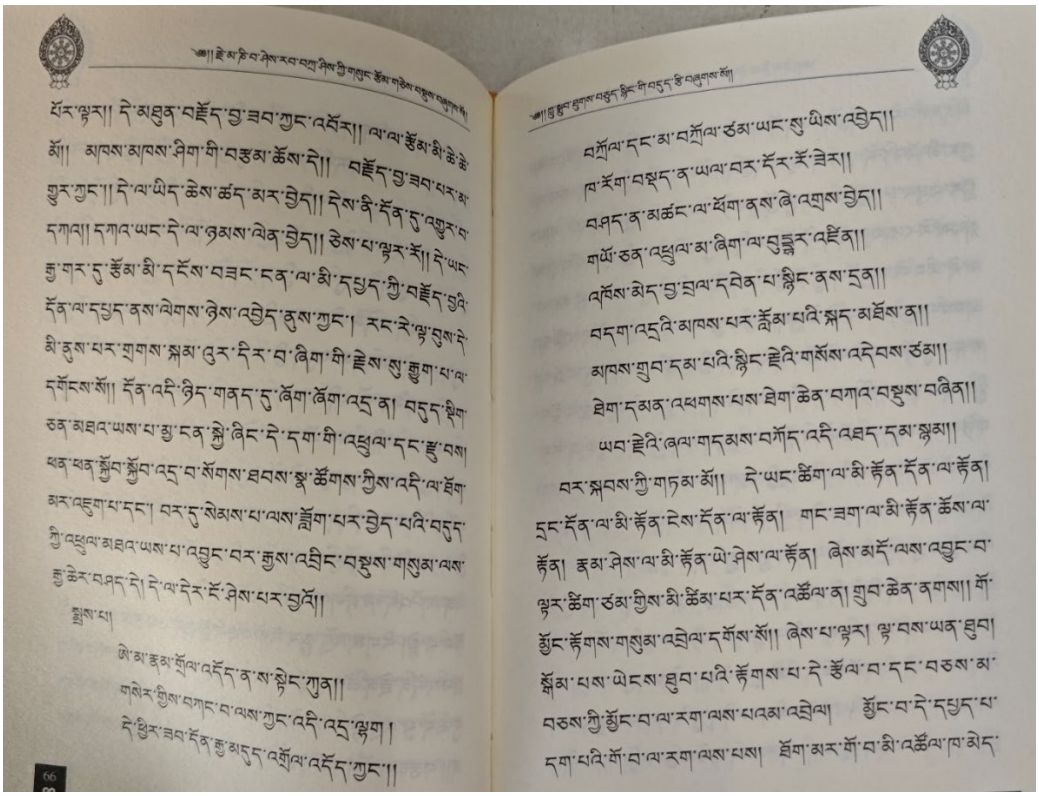

Fig. 7 Page 66 and 67 of [Suiduo, 1999], showing ||

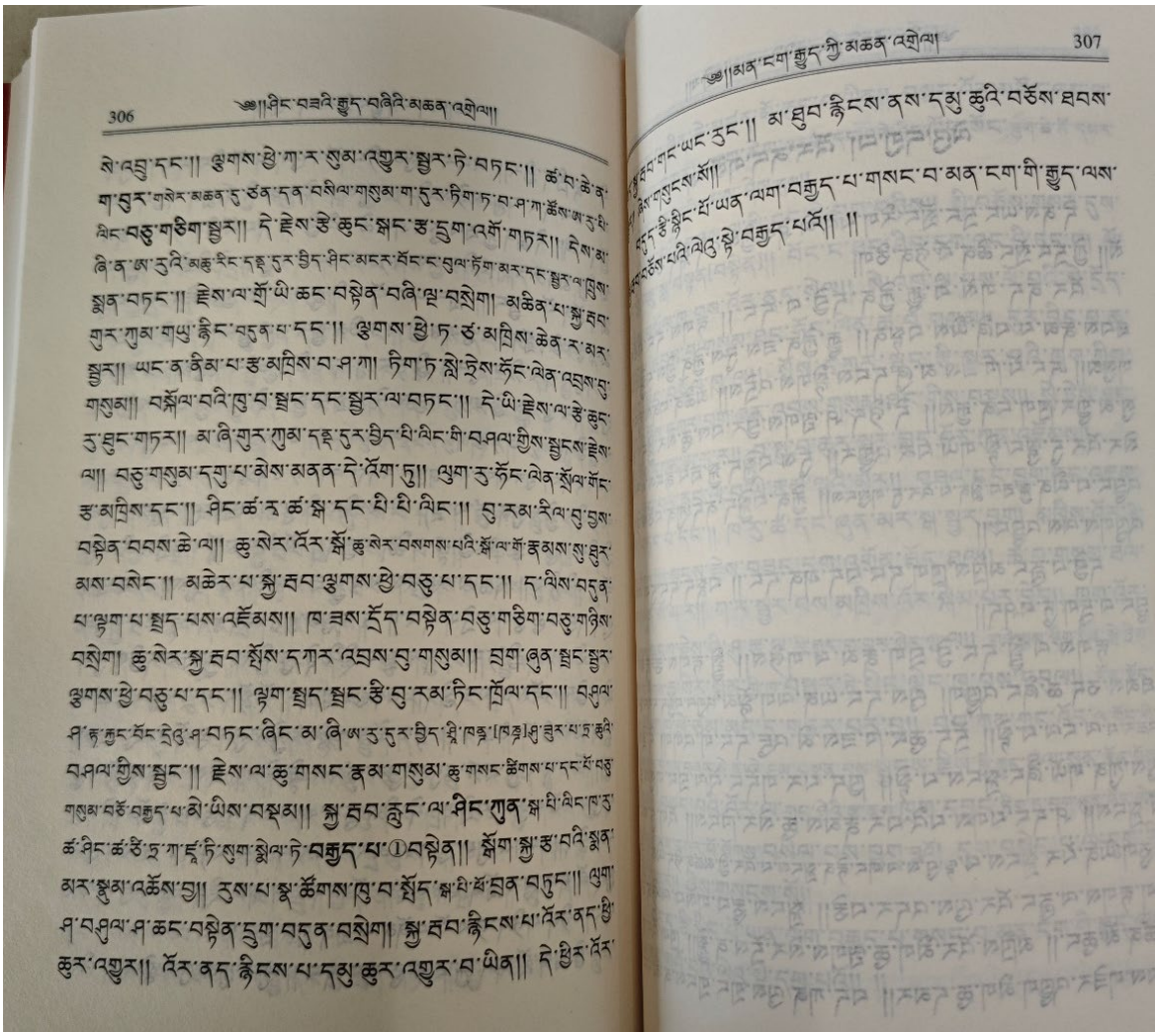

Fig. 8 Page 306 and 307 of [Shingbzav, 2018], showing || and || ||

Some important information can be found in the example above, especially **Fig. 5** and **Fig. 6**.

First, when there is extra space between two *shads*, the exact size of that space is involved in the calculation of the equalization of the remaining white space in the row. See **Fig. 6** to note that the distance between the two *shads* in each line is different, but is greater than the distance between the two *shads* at the end of each paragraph.

Second, when a sentence ends with ཀ TIBETAN LETTER KA, ག TIBETAN LETTER GA or ཤ TIBETAN LETTER SHA, a *shad* needs to be omitted from the subsequent ending indicator as these letters are considered to carry a *shad*. See **Fig. 7** and **Fig. 8**, where *ག།། is replaced by ག།.

Third, when a sentence ends with ང TIBETAN LETTER NGA, a non-breaking tsheg ་ TIBETAN MARK DELIMITER TSHEG BSTAR needs to be inserted before the ending indicator. Examples of this are found in **Fig. 5** through **Fig. 8**. The second the third rules are described in Dawapengcuo et al. (2005).

Fourth, even though there is extra space between the two *shads*, it seems to produce a line break only after the second *shad*. See **Fig. 6**, and this rule is also found in some other historical and modern literature.

In summary, (1) since other decorative symbols may be inserted in the middle of the two *shads* that serve as the ending indicator, the two *shads* need not be considered as a whole character. (2) Since the spacing between the two *shads* is involved in the calculation of the remaining white space in the line, this space should not be determined in the type design phase; (3) Since the ending indicator will be required to omit a *shad*, the space between the *shad* that the letter carries and the other *shad* cannot be provided by either the letter or the *shad*, and thus the space should correspond to an additional character that is different from the two characters.

Therefore, two *shads* with spacing in the middle should be represented as ⟨U+0F0D ། TIBETAN MARK SHAD, U+00A0 NO-BREAK SPACE, U+0F0D ། TIBETAN MARK SHAD⟩, since it seems to produce a line break only after the second *shad*. If it is required to be able to produce a line break between two *shads*, it should be represented as ⟨U+0F0D ། TIBETAN MARK SHAD, U+0020 SPACE, U+0F0D ། TIBETAN MARK SHAD⟩. When a *shad* in the ending indicator is omitted, the corresponding string is modified to ⟨U+00A0 NO-BREAK SPACE, U+0F0D ། TIBETAN MARK SHAD⟩ or ⟨U+0020 SPACE, U+0F0D ། TIBETAN MARK SHAD⟩.

Since the spacing between two *shads* should not be determined in the type design phase, but should be involved in the calculation in the typesetting phase, the only requirement that can be made for U+0F0E ། TIBETAN MARK NYIS SHAD is that it be equivalent to ⟨U+0F0D ། TIBETAN MARK SHAD, U+0F0D ། TIBETAN MARK SHAD⟩. We also consider further deprecating U+0F0E, as described in Long et al. (2016).

In addition, we examine the China national standards for U+0F0D and U+0F0E. According to the China national standard GB 16959—1997 (*Information technology — Tibetan coded character sets for information interchange — Basic set*), the number 014 representative glyph corresponds to U+0F0E and is designed without an extra space between the two *shads*. For the China national standard GB/T 29276—2012 (*Information technology — Tibetan ideogram coded character set (basic set & extension set A) — 24×48 dot matrix font — Bkav bstan bold*) as an example, there is also no extra space between the two *shads* in the glyph image corresponding to U+0F0E, and no extra glyph image has been designed for U+0F0E. Compliance with these national standards would result in the behavior of U+0F0E being equivalent to ⟨U+0F0D, U+0F0D⟩ at the level of typography in general.

2 Proposal

The original text of the relevant paragraph in the *Core Specification* is:

> 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* are used at the end of whole topics (“don-tshan”). 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.

We will require it to be changed to (unsure about the last sentence of the addition):

> 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* are used at the end of whole topics (“don-tshan”). ~~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.~~ Each *shay* is represented as a U+0F0D, except when the expression or topic ends with the letters U+0F40, U+0F42, or U+0F64, the *shay* is omitted and does not correspond to a character. There may or may not be extra space between two *shays*, and when there is extra space between two *shays*, it is necessary to insert U+00A0 or U+0020 between the two U+0F0D, if a line break is expected after the second *shay* or after the first *shay*. U+0F0E was encoded to support the case of two *shays* that are not omitted and have no extra space in between, and it is equivalent to two U+0F0D. Starting with Unicode 17.0, this character is deprecated in favor of two U+0F0D.

Reference

Xi Jinping 习近平 (author), China Ethnic Languages Translation Center 中国民族语文翻译局 (translator) (2015a). ཞི་ཅིན་ཕིང་གིས་རྒྱལ་སྤྱིའུ་ལུགས་སྐོར་བཤད་བ། པོད་དང་པོ། / 习近平谈治国理政 第一卷. Beijing / 北京: Ethnic Publishing House / 民族出版社.

Xi Jinping 习近平 (author), China Ethnic Languages Translation Center 中国民族语文翻译局 (translator) (2015b). ཞི་ཅིན་ཕིང་། རྒྱལ་སྤྱིའུ་ལུགས་སྐོར་བཤད་བ། / 习近平谈治国理政. Beijing / 北京: Ethnic Publishing House / 民族出版社.

Caiwanglamu 才旺拉姆, Xu Lihua 徐丽华 (2021). *Research on the Tibetan Script* / 藏文研究. Beijing / 北京: Ethnic Publishing House / 民族出版社.

སྡས་རྡོ། 岁多 (ed.) (1999). རྡེ་མ་ཀེ་བ་ཤེས་རབ་བཀྲ་ཤིས་ཀྱི་གསུང་རྩོམ་གཅེས་བསྡུས་བཞུགས་སོ། / 嘛呢巴·喜饶扎西文选. Lanzhou / 兰州: Gansu Minzu Publishing House / 甘肃民族出版社.

ཤིང་བཟའ་སྐལ་བཟང་ཆོས་ཀྱི་རྒྱལ་མཚན། (author), མཚོ་སྨོན་ཞིང་ཆེན་བོད་ཀྱི་གསོ་རིག་ཞིབ་འཇུག་ཁང་། 《བོད་ཀྱི་གསོ་བ་རིག་པའི་གནའ་དུའི་སྒོགས་བསྡིགས་དའི་ཚོགས་》རྩོམ་སྡིག་ཚོགས་པ། (ed.) (2018). ཤིང་བཟའི་རྒྱུད་ཞིའི་མཚན་འགྲུལ། སྤྱི་ཚ། Beijing / 北京: Ethnic Publishing House / 民族出版社.

Dawapengcuo 达哇彭措, Ga Zangcao 尕藏草 (2005). “网络媒体中藏文版式规则”. In: 西北民族大学学报（自然科学版), (01): pp. 64–65+95. DOI: 10.14084/j.cnki.cn62-1188/n.2005.01.016.

Long Congjun 龙从军, Liu Huidan 刘汇丹, An Bo 安波, Cai Hua 才华, Wu Jian 吴健 (2016). “藏文编码字符集标准应用中的问题及对策”. In: 信息技术与标准化, (Z1): pp. 46–51.

(End of Document)