Extended Wylie Transliteration Scheme

By Nathaniel Garson & David Germano
Tibetan & Himalayan Digital Library

Introduction

> In view of the increasing interest in Tibetan studies, it is desirable now more than ever that serious consideration be given to the acceptance of a standard system of Tibetan transcription. It is time to trade transcriptional independence for uniformity in order to facilitate and standardize the advancement of Tibetan studies.[i]

These are the words of Turrell Wylie, who in 1959 proposed what is arguably the most widely used Tibetan transliteration system to date. With the recent burgeoning of digital technology in humanities and the surge of scholarly interest in Tibetan studies, the need for a standard transliteration system is even more compelling at present. However, Wylie's scheme is not entirely comprehensive, and incommensurable extensions of it have proliferated in the digital world. While the international standards community ("ISO") has established standards for the transliteration of other languages using non-Roman alphabets, no such standard has been established for Tibetan, either de facto or ISO-approved.[ii] We thus feel the time has come to establish an initiative in order to promote a standard format for information exchange dealing with Tibetan texts and language.

Wylie's fundamental argument was that a Tibetan transcription system "should be of minimal complexity and capable of reproduction on a standard typewriter, i.e. one lacking special keys for diacritical marks." Implicit in his method is an attempt to provide rough phonetic equivalents for the transliterations of each Tibetan letter. His system

---

[i] Turrell Wylie, "A Standard System of Tibetan Transcription," *Harvard Journal of Asiatic Studies*, 22 (1959), 263.

[ii] The International Organization for Standardization (ISO) has implemented standard transliteration schemes for a number of different languages. For instance, the transliteration standard for Thai is ISO 11940:1998.

succeeds in avoiding the use of diacritical marks and the anomalies employed by other systems and has been widely adopted by many scholars since its introduction.

For this reason, it seems clear that Wylie's scheme should be the basis for any proposed standard but that certain deficiencies need to be addressed in order for the system to be useful in a digital environment. This document presents a proposal for such an extended Wylie transliteration scheme. Our system is geared towards disambiguated transliteration, which is required for digital processing and aims to change as little as possible the already extant scholarly practices in most countries, especially those where transliteration is most commonly used with respect to Tibetan. It therefore does not alter the basic equivalents set forth by Wylie, as these have gained wide acceptance and are logically consistent. Though some Tibetologists in the People's Republic of China have argued that the a-chung and a-chen should be represented by 'v' and 'x' because they are treated as consonants in most Tibetan grammars, we see no compelling reason for such an alteration. First, the apostrophe used to represent a-chung in Wylie is not a vowel; second, the status of a-chung and a-chen in Tibetan grammar is controversial; and third, while Wylie is not a phonetic system, it does have the advantage of using reasonable phonetic approximations of Tibetan letters; the 'v' and 'x' would completely contradict that principle.

Though we have adopted Wylie's basic system, there are several fundamental gaps in his transliteration scheme that need to be addressed, if one is to develop a comprehensive standard. These are:

1. The inability to represent Tibetan transliterations of Sanskrit characters not normally found in Tibetan language *without the use of diacritics*,
2. The inability to *unambiguously* represent unusual stacks of Tibetan characters in the transliteration of mantras and the like, and
3. The lack of equivalents *for the full range* of Tibetan punctuation marks and a few new Tibetan characters used to represent non-Tibetan vocalizations, such as ཪྻ and ཫྻ.

Various solutions have been proposed to these problems, though there has been no broad consensus. The following presents an unambiguous and comprehensive solution

to these problems, with an eye toward digitally processing Tibetan documents as well as orthographic considerations. *This is a provisional proposal*, which we plan to systematically revise as necessary in coordination with critical feedback from Tibetan scholars and Tibetan computing experts from around the world. Our goal is to establish an internationally accepted standard that is both used by the scholarly community *and* approved by the relevant international standards organization. Wylie's basic scheme in regards to transliterating the consonant and vowels is straightforward and needs no modification or explanation. It is outlined in tables I and II below.[iii] The remainder of this paper explains the proposed additions and modifications to Wylie's system.

In this technology age, a transliteration scheme must serve the needs of both print publication and electronic publications. Print publications most typically use transliteration for references to terms or names, as well as for citations of short passages; in critical editions, an entire text may be given in transliteration rather than Tibetan script. The use of transliteration in print publications circumvents the difficulties of editing, proofing, and typesetting Tibetan script, as well as being at least somewhat accessible to those who do not read Tibetan script. For electronic publications, the same concerns apply. Moreover, in the digital medium, a standard transliteration scheme would also provide enhanced functionality. First, until Tibetan Unicode is fully implemented, digital text processing of Tibetan – searching for keywords, sorting, and the like – is easier with Roman transliteration, because of the present state of operating systems and software. Second, data stored as transliteration is capable of being transformed and displayed in different forms—Tibetan script, phonetic rendering, or even another transliteration system. This is only possible if there is an *absence of ambiguity* in the transliteration system so that any combination of letters has a single interpretation that cannot be confused with another ligature. The Extended Wylie Transliteration Scheme proposed here outlines a disambiguated system that uses standard letters and symbols found on the average computer keyboard to cover the full range of possible Tibetan ligatures. Tibetan transliteration of Sanskrit is represented

[iii] There is a difference in how Extended Wylie deals with Sanskrit-related vowels in Chart II. This will be discussed anon.

using related capital letters and the plus sign, while unusual Tibetan punctuation marks are assigned symbols commonly found on the computer keyboard, such as the percent sign (%), the exclamation (!) and so forth.

One potential problem to this solution is that many authors have taken to using capital letters to distinguish proper nouns. Wylie's argument against this practice, back in 1959, speaks for itself:

> The phonetic argument maintains that the initial should be capitalized as an indication to pronunciation and in order to distinguish prefix letters, which are silent in the dialect of Central Tibet. The application to English of such "phonetic" capitalization would result in such spellings as: hOur, kNight, pNeumonia, pSychiatry, and phTisic. The phonetic argument suffers from the same practices as the lexicographic argument in that only the first word of a textual passage or of compound names is subject to capitalization. This evidences the assumption that the reader is familiar with Tibetan pronunciation; and if this be the case, what need is there for phonetic capitalization?[iv]

One possible compromise to these conflicting methods is to allow "phonetic" capitalization for Tibetan proper nouns embedded in non-Tibetan text, such as English, but to adhere to extended Wylie transliteration practices within formal citations.

Tibetan Transliteration of Sanskrit

Because of the great importance of Indian Buddhist scriptures for Tibetan religious culture, Tibetans from the eighth century CE onwards developed a method for accurately transliterating Sanskrit using Tibetan orthography. However, to do so they had to make modifications to their glyphs and writing conventions. There were two basic problems. First, Sanskrit contains a whole series of characters not found in Tibetan, mainly the retroflex characters but also such characters as the *anusvara* and *visarga*, to name a few. These last two were dealt with by adopting the Sanskrit ligature—a circle above the

[iv] Wylie, "A Standard System of Tibetan Transcription", 265-266.

character and a colon-like glyph at the end of a word, respectively. For the retroflex characters, they took their closest relative, generally the dentals with one exception, and reversed them. Similarly, the Sanskrit semi-vowels, ṛ and ḷ, are formed by a combination of the appropriate consonant in its subscribed form and a reversed *gi gu* or "i". Wylie's system presents no way to transliterate the Sanskrit retroflex consonants, semi-vowels and long vowels without resorting to diacriticals, thereby transgressing Wylie's own guidelines.

The second problem was that Sanskrit contains consonant combinations that do not conventionally occur in Tibetan. The Tibetans solved this problem by stacking the consonants with the first on top and the last on the bottom, though such combinations often transgressed the traditional grammatical rules for creating stacks in Tibetan. Wylie's system can account for these abnormal stacks by stringing the consonants together. However, this can lead to ambiguities in the transliteration, which would present serious problems to digital processing.

In the system proposed here, retroflex characters are handled in a way similar to the Tibetan method, namely by capitalization of the corresponding dental. There are five main retroflex characters, represented below with their diacritical transliteration:

ཊ་ ṭa ཋ་ ṭha ཌ་ ḍa ཎ་ ṇa ཥ་ ṣha

Even with the computer, diacritical marks are hard to enter, requiring several keystrokes for one character. Yet, if a simple, unambiguous transliteration is adopted the computer *can* easily be programmed to represent those characters with diacritical marks if so desired, as well as correspondingly represent them with the correct Tibetan glyph. Thus, it is proposed that for these characters, the capital forms of their corresponding letters be used, rendering them:

Ta Tha Da Na Sha

Similarly, the Sanskrit semi-vowels, ṛ and ḷ, would be also handled through capitalization, becoming "Ri" and "Li" respectively; their long forms would be accounted for by capitalizing the "i", producing "RI" and "LI". Finally, the long vowels, which in Tibetan are transliterated by subjoining an *a-chung* or འ to the vowel-stack combination, are also handled through capitalizing the main vowel.

It makes sense to use capitalization for the less common Sanskrit-derived characters, as capitalization is not the norm in Tibetan and requires an extra keystroke. This would of course preclude using capitals to distinguish proper names or root letters. As has been shown, there is little practical benefit in distinguishing root letters that way. Capitalization of proper names could be allowed when Tibetan is imbedded in English for conformity to English practice, but should not be used for strict transliteration. For the sake of consistence it is proposed that the *anusvara* and *visarga* be handled in a similar way through capitalization:

M H

While the objection could be raised that interspersed capital letters are difficult to read, the purpose here is to develop an *unambiguous* and *simple* transliteration system as Wylie proposed. In print publications, diacritics could still be used, as they are today. However, with digital technology, it is desirable to have the transliterated text easily transferable between machines and platforms. The use of diacritics, which are font specific, hinders this, whereas the use of capitals does not. Conversely, conversion from our extended Wylie system to a diacritic font would be relatively straightforward, as long as the transliteration system is, like this one, disambiguated.

Non-standard Tibetan stacks found in Tibetan transliteration of Sanskrit also present a problem for transliteration. These stacks can be represented in Wylie's system by merely stringing consonants together. However, some of these Sanskrit stacks can be confused for letter combinations found in standard Tibetan. For instance, there is a Sanskrit stack consisting of an *n* over a *y.* Unless provision is made for distinguishing this combination, it would naturally be confused with the eighth Tibetan letter, *nya*. For this situation, we suggest the use of the plus sign (+) between all non-standard stack letters found in Tibetan transliteration of Sanskrit. For instance, there are aspirated letters in Sanskrit not found in Tibetan, such as *dha*, *gha*, and *bha*. These would be transliterated as *d+ha, g+ha,* and *b+ha* respectively.

We have however retained the use of the period between a *g* prefixed to a *y* to distinguish such a combination from the stack *gy*, because this is an already well-established practice and is to prevent stacking. Thus, the Tibetan words for right and left are transliterated *g.yon* and *g.yas*, whereas the word for ornament would be *rgyan*.

Punctuation

The other major omission of the Wylie scheme lies in its limited coverage of punctuation. This system adopts the basic policy, as does Wylie, of using a space for a *tsheg* (་), the intersyllable dot, with the special exception of using an asterisk (*) for a non-breaking *tsheg*, which will not allow a line break to occur after it. It also conforms to the common convention of using the slash (/) for a *shad* (།), the vertical line ending a phrase. In other Wylie-based transliteration systems that also use a space for a *tsheg*, there is no way to represent the white space that can actually occur in Tibetan texts. To disambiguate this situation, the underscore (_) is proposed as an equivalent to Tibetan white space.[v] Besides the *tsheg* and *shay*, Wylie presents no policy concerning punctuation in his article. Yet, Tibetan has as much punctuation as English, if not more. It is therefore necessary to develop a standard for transcribing these unpronounced marks. The remaining punctuation marks are less frequent, though the attempt has been made to represent them with standard keyboard strokes whose characters are if possible somewhat similar in form to the corresponding Tibetan glyph. Thus, the *tsheg shad* (།), a line with a dot above it, is represented by the semicolon (;). The *rin chen spungs shad* (།), a *shad* with three dots above it, is represented by the vertical bar (|). The *sbrul shad* (།) is represented by an exclamation mark (!), and the *gter shad* (།) represented by the colon (:). Though it could be argued, that the colon is more appropriate for the *visarga*, as they are almost identical in shape, the counter argument is that the *gter shad* not only is similar in form to a colon but they are both punctuation marks (albeit with different functions), whereas the *visarga* represents an aspiration at the end of the word and given the method of dealing with transliterated Sanskrit above, is best represented by a capital H. The final punctuation mark to mention is the *zla tse* (), which has two forms—the full, single form () and the truncated form () appended after the full form to form, . The full form is represented by a similar

[v] It would be up to software manufactures and encoders to determine whether the whitespace character had a fixed or variable size, depending on the degree of accuracy desired in representing Tibetan texts digitally.

looking glyph, the at-sign (@), while the truncated form is represented by the next symbol over on the keyboard, the pound sign (#) for ease of typing. This covers the major punctuation marks; others are listed in the punctuation chart below.

Not all the punctuation marks in the chart below have transliteration equivalents. This is because there are not enough characters on the standard keyboard to correspond to every character that is found in the Unicode 3.0 Tibetan character set. Besides, there should be a secondary way of transcribing a character as some keyboards may lack certain keys such as the dollar-sign and so forth. To this end, it is proposed that the use of an "escape" character is standardized. The "escape" character would be inserted before the desired character's Unicode value, just as one can now use the alt key to type in a decimal code for a character. This can be done with any character in the Tibetan character set. The proposed escape character is the backslash (\) and the code should be the Unicode 3.0 Tibetan character code. Thus, the *rgya gram shad,* ( – transliterated by a left bracket:]) could also be encoded by typing a back slash followed by its Unicode hexadecimal equivalent: \0F12. The astrological sign *sgra gcan char rtags* (), lacking a transliteration equivalent, will have to be entered: \0F17. As the Unicode is unambiguous and always four characters long, there is no need for a closing marker. Because the consonants, vowels, Sanskrit-derived characters, and punctuation make use of every available key on the standard keyboard, such escape codes will have to be used for the other Tibetan characters in Unicode 3.0 (Chart VI) and later releases. These escape codes could be used for other character sets as well, such as Devanagari or Chinese, which may be interspersed with Tibetan. This accounts for all the Unicode 3.0 character set; however, some provision needs to be made to deal with punctuation not included in that set. In the last chart below (Chart VII), we have included some marks found in Tibetan Computer Company's Tibetan Machine font that do not appear to be included in the Unicode standard for Tibetan.

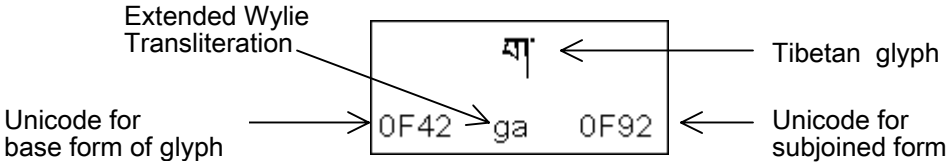
The Extended Wylie Transliteration Scheme thus proposed covers all the various possible letter combinations found in Tibetan literature in an unambiguous way. For most situations, Wylie's basic system will suffice. So, it has been left in tact here. However, if the more unusual letter combinations are found, they can be easily and clearly transliterated using the above system. The following charts give the complete

proposed Extended Wylie system of transliteration as described above. There are seven tables:

I. Consonants,
II. Vowels,
III. Numerals,
IV. Sanskrit letters,
V. Punctuation,
VI. Other Tibetan characters found in Unicode 3.0,
VII. Characters not found in Unicode 3.0.

The transliteration of all characters, except the vowels, includes the short *a* sound implicit in all Tibetan characters. This sound is only pronounced when that character is the root character of the syllable (*tsheg bar*), but not when it is a prefix, superscript, subscript, or suffix. The same is true in transliteration, such letters dropping their implicit *a.* The vowels are displayed here in their common stand-alone form using the *a-chen* as their base, though they of course can be appended to any root letter or stack.[vi] As with other Sanskrit derivatives, the long simple vowels are represented by capitals, while complex vowels are represented by the corresponding diphthong.

For each character in each of the charts, the Tibetan glyph is given in the middle of the cell, the proposed transliteration is given immediately below it in boldface. To the left of that is the main Unicode character code, and to the right, if there is a number, is the Unicode character code for the alternate subscribed version of that character.



[vi] Except in the case of the Sanskrit semi-vowels, *Ri/RI* and *Li/LI*. They are given in combination with the consonants *ha* and *ka* respectively with which they are in more natural combination.

I. Consonants

	ཀ་			ཁ་			ག་			ང་	
0F40	ka	0F90	0F41	kha	0F91	0F42	ga	0F92	0F44	nga	0F94
	ཅ་			ཆ་			ཇ་			ཉ་	
0F45	ca	0F95	0F46	cha	0F96	0F47	ja	0F97	0F49	nya	0F99
	ཏ་			ཐ་			ད་			ན་	
0F4F	ta	0F9F	0F50	tha	0FA0	0F51	da	0FA1	0F53	na	0FA3
	པ་			ཕ་			བ་			མ་	
0F54	pa	0FA4	0F55	pha	0FA5	0F56	ba	0FA6	0F58	ma	0FA8
	ཙ་			ཚ་			ཛ་			ཝ་	
0F59	tsa	0FA9	0F5A	tsha	0FAA	0F5B	dza	0FAB	0F5D	wa	0FAD[vii]
	ཞ་			ཟ་			འ་			ཡ་	
0F5E	zha	0FAE	0F5F	za	0FAF	0F60	'a	0F71[viii]	0F61	ya	0FB1[ix]
	ར་			ལ་			ཤ་			ས་	
0F62	ra	0FB2[x]	0F63	la	0FB3	0F64	sha	0FB4	0F66	sa	0FB6
	ཧ་			ཨ་							
0F67	ha	0FB7	0F68	a	0FB8						

[vii] 0FAD is the Unicode for the common *wa-zur* subjoined character, which has its own abbreviated glyph. A separate code, 0FBA, is reserved for the less common fully formed, subjoined *wa.*

[viii] 0F71 is the Unicode for the common *'a-chung* subjoined character, which is written smaller. A separate code, 0FB0, is reserved for the less common fully formed, subjoined *'a.*

[ix] OFB1 is the Unicode of the common *ya-btags* character, which has its own abbreviated glyph. A separate code, 0FBB, is reserved for the less common fully formed, subjoined *ya*.

[x] OFB2 is the Unicode of the common *ra-btags* character, which has its own abbreviated glyph. A separate code, 0FBC, is reserved for the less common fully formed, subjoined *ra*.

II. Vowels

ི 0F72 i	ུ 0F74 u	ེ 0F7A e	ོ 0F7C o
ཱི 0F73 I	ཱུ 0F75 U	ཻ 0F7B ai	ཽ 0F7D au
ཱ 0F71 A	ྲི 0F76 (h)Ri	ཱྲི 0F77 (h)RI	ླི 0F78 (k)Li
ཱླི 0F79 (k)LI	ྀ 0F80[xi] -i	ཷ 0F81 -I	

III. Numerals

๐ 0F20 0	๑ 0F21 1	๒ 0F22 2	๓ 0F23 3	๔ 0F24 4
๕ 0F25 5	๖ 0F26 6	๗ 0F27 7	๘ 0F28 8	๙ 0F29 9

[xi] This character is called TIBETAN VOWEL SIGN REVERSED I. Its transliteration is problematic because the *gi gu* to which it is related has already been assigned "i". However, as the phenomenon of reversing characters does in fact occasionally occur elsewhere in Tibetan, the proposal here is to use a general method for indicating such reversal. Just as the plus-sign (+) will be used to indicate a non-standard Tibetan stack, the minus sign (-) will be used for non-standard reverse letters, so that a reversed *ha* would be *-ha*. The only exception would be the commonly found reversed dentals (*ta*, etc.), representing the Sanskrit retroflex letters. These will be represented by capital letters. The question then arises should the reverse *d* that is an abbreviation for *gs* at the end of *tshig bar* be encoded differently from the retroflex reversed *d*. If so, the former could be *-d* while the latter *D*.

IV. Sanskrit Letters

A. Aspirates and the Compound *k+Sha*

ཀྷ 0F43 g+ha 0F93	ཌྷ 0F52 d+ha 0FA2	བྷ 0F57 b+ha 0FA7	ཛྷ 0F5C dz+ha 0FAC
ཀྵ 0F69 k+Sha 0FB9			

B. Retroflex

ཊ 0F4A Ta 0F9A	ཋ 0F4B Tha 0F9B	ཌ 0F4C Da 0F9C	ཎ 0F4D D+ha 0F9D
ཎ 0F4E Na 0F9E	ཥ 0F69 Sha 0FB9		

C. Other Sanskrit-related Letters

ཿ 0F7F H	ཾ 0F7E M	ཱཾ M^	ཱ 0F83 ~	ཱྀ 0F82 ~^
ྀ 0F84 ? 0F9E	ཱྃ 0F85 & 0FB9			

V. Punctuation

0F0B (space)	0F0C *	(*Whitespace*)[xii] — (*underscore*)	 0F0D /	0F0F ;
0F11 |	0F08 !	0F14 :	0F10 [	0F12]
0F13 `	0F04 @	0F05 #	0F06 $	0F07 %
0F34 =	0F3A <	0F3B >	0F3C (	0F3D)
0F3E {	0F3F }			

VI. Other Tibetan Characters in Unicode 3.0

0F00 oM	0F01	0F03	0F03	0F09
0F0A	0F15	0F16	0F17	0F18
0F19	0F1A	0F1B	0F1C	0F1D

[xii] These first two "punctuation marks" need some explaining. The first is the *tsheg*, a small dot written between syllables or semantic units in Tibetan, which themselves are called *tsheg bar* ("between *tsheg*s"). In Wylie transliteration, the *tsheg bar* were separated by spaces. This has become the standard and will be used here. There are two kinds of *tsheg* listed in Unicode 3.0. The second is a non-breaking space (0F0C), whose equivalent will be an asterisk. However, Tibetan does include whitespace between its phrases and sentences. So as to eliminate ambiguity, the whitespace needs to be represented by a different character, namely the underscore.

0F1E	 0F1F	 0F2A	 0F2B	0F2C
 0F2D	0F2E	 0F2F	0F30	0F31
 0F32	0F33	 0F35[xiii]	0F36	0F37[xiii]
0F38	0F39	 0F86	0F87	0F88
 0F89	0F8A	 0F8B	0FBE	0FBF
0FC0	0FC1	 0FC2	0FC3	0FC4
 0FC5	0FC6	 0FC7	0FC8	0FC9
0FCA	0FCB	0FCC		

VII. Tibetan Characters not in Unicode 3.0

[xiii] These two glyphs are similar to the *bindu* and *chandra bindu*. However, the characters differ in that they are placed below a syllable and are used to highlight a person's name.