Title:	Comments on N2665, Opinions on Encoding Tai Lue
Doc. Type:	Expert contribution
Source:	Peter Constable, Microsoft
Date:	October 16, 2003
Action:	For consideration by JTC1/SC2/WG2, UTC
References:	WG2/N2634 (=L2/03-321), N2242R, L2/99-243
Distribution:	WG2 members, UTC members

Document N2634 provides a complete proposal for encoding the Tai Lue (Xishuangbanna Dai) script. N26xx is a contribution from the China national body providing comments on that proposal. This document provides comments on the Chinese contribution.

Comments on individual points

The Chinese comments are listed in eight numbered points. The following comments are numbered to correspond with the points in the Chinese contribution.

1. It is noted that the script in question is mainly used in China. This is undisputable. The Chinese body hopes that N2242R, submitted by China at WG2 meeting #39, will be recognized as the original proposal. There were, in fact, earlier proposals or draft proposals (as is noted in N2242R itself). What is important to determine, however, is not which proposal was first, but what proposal is the best for encoding the script. Comparison of the two proposals, N2242R and N2634, is provided below.

2. It is noted that the number of characters in the two proposals is not the same. The is due primarily to a difference in encoding model: N2242R analyzes certain text elements into base + combining mark sequences while N2634 treats these as atomic characters (and does not propose separate combining marks). The only difference in terms of representable text is that N2242R omits the digit zero, which is attested in Tai Lue documents, as shown in Figure 1:[1]

[1] This sample illustrates the existence of two alternate typeforms for the digit 1, both of which were listed in L2/99-243.

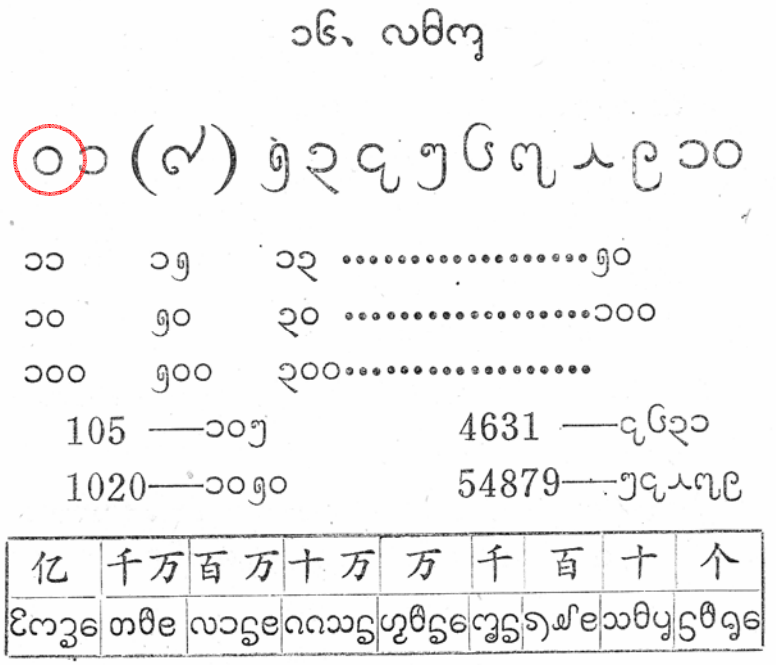

Figure 1. Tai Lue numbers 0—10 (LA1, p. 39)

Differences between N2242R and N2634 will be discussed further below.

3. It is noted that the typeface used in N2634 does not conform to the conventions and way of writing of the Tai Lue people. In fact, there are significant variations in typefaces, as is demonstrated from samples taken from [LA1] and Jinxiang 1994:

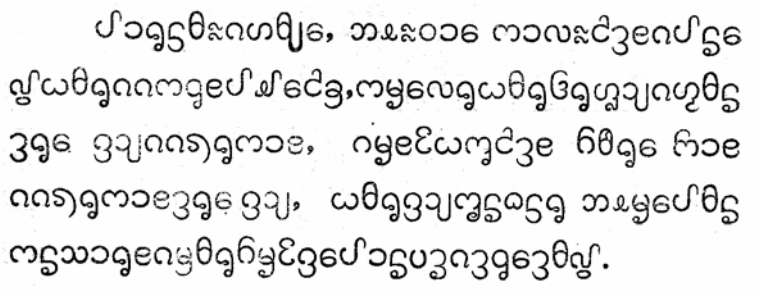

Figure 2. Tai Lue type sample: LA1 (p. 73)

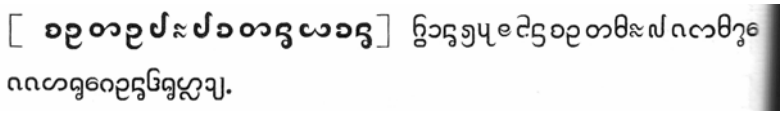

Figure 3. Tai Lue type sample: Jinxiang 1994 (p. 73)

Figure 4. Tai Lue type sample: Jinxiang 1994 (cover)

Jinxiang itself uses two very distinct faces: the cover uses a display face with rounded typeforms, as are seen in the body text from [LA1], in contrast with the squared typeforms of the body text in Jinxiang 1994. The body face in Jinxiang has significant thick/thin variation in stroke widths, whereas the body face of [LA1] has fairly uniform stroke widths.

The typeface used in the code chart of N2634 was developed in the USA, but was created for use in projects of the Chinese Academy of Social Sciences, Institute for Nationality Studies. The design was intended to be original and to add qualities to aid in legibility, but was guided strongly by the typeface used in [LA1]. It is the only digital typeface for Tai Lue script that I know of, though I see no objection to use of a different face for published code charts, if another and better one is available.

4. The description of spelling rules in section E of N2634 is said to be unclear. It seems to me that the intent is made clear by the detailed examples in Table A of N2634: orthographic syllables can include a consonant, vowel and tone mark, and the tone mark is written following the consonant and any vowel marks.

The authors of N26xx inquire what the meaning is of the statement, “which are modified with a hook showing that the inherent vowel is killed”. This refers to the seven syllable-final consonant characters, which are common to both proposals: “ᧁᧂᧃᧄᧅᧆᧇ”.

The authors of N26xx inquire about the meaning of the statement, “Tai Lue consonants denote two tonal registers.” As is the case with other Tai scripts, such as Thai, Lao and Lanna, characters for obstruent phonemes come in pairs: two characters have the same segmental value, but differ with respect to the tonal values of syllables in with they are the initial consonant. For instance, the Tai Lue characters “ᦂ” and “ᦅ” both represent voiceless velar stop phonemes; in syllables with no written tone mark, syllables beginning with “ᦂ” have a high tone, whereas syllables beginning with “ᦅ” have a falling tone. Such pairings occur for all obstruent characters, and the two sets are referred to as “high class” and “low class” respectively.[2] These facts are generally very familiar to linguists studying Tai languages of China, Thailand, Laos, Myanmar and Vietnam.

5. It is stated that the examples in Table A of N2634 are invalid syllables in Tai Lue, and that some of the phonetic transcriptions are wrong. There is no explanation as to what precisely is incorrect in Table A. These criticisms should not have a bearing on the character encoding proposal of N2634, however: all that is truly important from Table A is the illustration of the encoding model by means of example character sequences and the corresponding display.

[2] “High” and “low” are direct translations for the terms used in Tai languages to refer to these classes. For instance, these classes are called ᦉᦳᧂ /suŋ ˥/ ‘high’ and ᦶᦎᧄ /tɛm ˦˩/ ‘low’.

7. The authors of N26xx note that "Tai Lue scripts include New Tai Lue and Old Tai Lue," and that space should be provided in the BMP for both. It is not clear whether the authors intend that both should be included in the same block.

In fact, space has been allocated in the range U+1A80–U+1AFF for Old Tai Lue, or "Old Xishuangbanna Dai", script under the name "Lanna".[3] This is appropriate since the older tradition for writing Tai Lue uses a variant of the Lanna script, which originated in the Lanna kingdom of northern Thailand and was adopted for writing various Tai languages in surrounding regions, including Khun to the west, Tai Lue and perhaps Yong to the north, and northern varieties of Lao to the east. It would be inappropriate to merge the older and newer writing systems into a single block as these are very distinct scripts with different character inventories and significantly different behaviours.

8. It is noted that "'Dai' is an internationally standardized name in Pinyin for Dai nationality and Dai script," and on that basis it is requested that "Dai" rather than "Tai" be used when encoding the script in the ISO/IEC 10646. It is true that "Dai" is the Pinyin Romanization used in China for languages of this family and for the scripts associated with these languages. At the same time, "Tai" is the Romanization most widely used internationally by linguists specializing in this family of languages. Note, for instance, the practice of the highly-esteemed Chinese linguist, Fang Kuei Li, who consistently used "Tai" in his English-language publications.

I also note that the Dehong script was named "Tai Le" when encoded in ISO/IEC 10646.

That said, I do not feel this distinction in naming conventions is a serious issue, but concede that it may be so for the Chinese national body.

Differences between proposals in N2242R and N2634

The preceding comments address individual points in N26xx. As noted in the comment on point 1, however, what is most important is to evaluate which proposal provides the best approach to supporting the script. It is important, then, to identify substantive differences between the proposals in N2242R and N2634.

N2242R did not propose appropriate names for characters, it did not explain encoding model or suggest character properties, nor did it provide documentation illustrating attested usage. This comparison will disregard those limitations and focus on substantive differences of the proposed character encodings.

As noted, N2242R omits the digit zero, which clearly appears to be an oversight. N2242R fails to provide certain key details, particularly whether vowel characters are considered combining marks, always encoded after the consonant, or whether they are spacing letters, preceding or following the consonant according to their visual order. I will assume the intent for N2242R was that they be considered combining marks.

The remaining differences between N2242R and N2634, then, are twofold: inventory of characters, and the ordering of characters.

[3] See the "Roadmap to the BMP," at http://www.unicode.org/roadmaps/bmp/.

Inventory of characters

The two proposals differ in character inventory due to a different encoding model with regard to certain text elements: N2242R uses a decomposed representation using combining marks, while N2634 uses combining marks. The text elements for which the encoded representations differ are listed in Table 1:

Text element	Representation in N2242R	Representation in N2634
ᦀ	ᦁ xx81 + ◌̂ xxC8	ᦀ 1981
ᦡ	ᦔ xx92 + ◌̂ xxC8	ᦡ 19A3
ᦢ	ᦗ xx93 + ◌̂ xxC8	ᦢ 19A5
ᦶ◌	ᦵ◌ xxA9 + ᦵ◌ xxA9	ᦶ◌ 19C2
ᦨ	ᦂ xx82 + ◌̥ xxC6	ᦨ 19A6
ᦪ	ᦅ xx94 + ◌̥ xxC6	ᦪ 19A7
ᦩ	ᦃ xx83 + ◌̥ xxC6	ᦩ 19A8
ᦫ	ᦆ xx95 + ◌̥ xxC6	ᦫ 19A9
᧟	᧞ xxC7 + ◌̥ xxC6	᧟ 19DF

Table 1. Different encoding models for 9 text elements

This difference is not huge, and either encoding could be implemented successfully. Overall, however, I find the advantages of N2634 sufficient to recommend its representation for these text elements over that proposed by N2242R.

The “hat” that occurs over the graphemes for the high class voiced stops, “ᦀᦡᦢ” does not combine productively with any other characters. By analyzing these graphemes into base + diacritic sequence, the pairing of characters in high class / low class pairs is broken, and various text processes are made slightly more complex as a result. For these reasons, the encoding of these three text elements proposed in N2634 is considered preferable.

By analyzing the vowel “ᦶ◌” into a sequence of two “ᦵ◌” characters, some text processes will be made slightly more complex in that the structure for encoded representation of orthographic syllables can now contain either one or two re-ordering vowels. Following the treatment of cognate vowel signs in other Indic scripts (Thai U+0E41, Lao U+0EC1, Sinhala U+0DDB and Malayalam U+0D48), it is considered best to encode this grapheme as a single character, as proposed in N2634.

The subjoined “va” sign occurs in Tai Lue text with four different consonants, “ᦂᦅᦃᦆ”, and with the abbreviation character “᧞”. It is not subjoined productively with any other consonants in Tai Lue. Analyzing it as a separate combining mark might be advantageous should the script ever be used for other languages that allow other labialized clusters. N2242R might be considered preferable over N2634 in this regard, therefore. In Tai Lue sources, the labialized forms are consistently listed as distinct graphemes. On the other hand, it is not a requirement that graphemes be encoded as atomic characters. It should also be noted that N2242R is slightly inconsistent in that the “va” sign is not analyzed in the high-class grapheme “ᦛ”. Also, I am not aware of any other languages that are written with this script and that require additional labialized forms. The use of combining marks in N2242R also implies the need for complex-rendering support in software implementations, though this is also needed with the proposal in N2634 to deal with reordering combining marks. In conclusion, I favour the treatment of N2634 slightly, but cannot give a strong recommendation of one treatment of labialized forms over the other.

Ordering of characters

There are differences of ordering between N2242R and N2634 in three different areas: (1) ordering of consonants, (2) ordering of diphthongs, and (3) ordering of syllable-final consonants in relation to vowels.

1. **Ordering of consonants.** Consonants within Indic scripts, including Tai scripts, are ordered according to several properties:[4]

   - point of articulation: velar, palatal, retroflex,[5] alveolar, labial
   - manner of articulation: e.g. voiceless unaspirated, voiceless aspirated, nasal
   - tone class: high, low

   The two proposals agree on the ordering within any one of these properties, as do all source materials I have encountered. The proposals differ, however, with regard to the priority these properties have in relation to one another. Specifically, they differ in whether tone class is given first priority or last priority. N2242R makes tone class the primary factor, and so has all high-class consonants preceding all low-class consonants. N2634, however, considers tone class last, and so interleaves high- and low-class consonants on the basis of pairs having common segmental values.

   The difficulty in determining the correct order is that source materials do not always make absolutely clear how tone class should be considered in relation to other consonant properties. It is not uncommon to display consonant charts with the high and low consonants listed in separate columns, but such charts do not make clear how tone should be considered in ordering, and so the division of tone classes in these charts should not necessarily be understood to imply anything with respect to collation. For instance, [LA1] contains such a chart, shown in Figure 5:

[4] Some collation conventions may divide manner of articulation into multiple factors; e.g., obstruent vs. nasal, and stop vs. fricative; or voiceless vs. voiced vs. nasal, and with or without aspiration.

[5] Tai languages do not have retroflex consonant phonemes, though some Tai scripts have maintained characters corresponding to retroflex consonants in Pali or Sanskrit. Tai Lue does not do so, however.

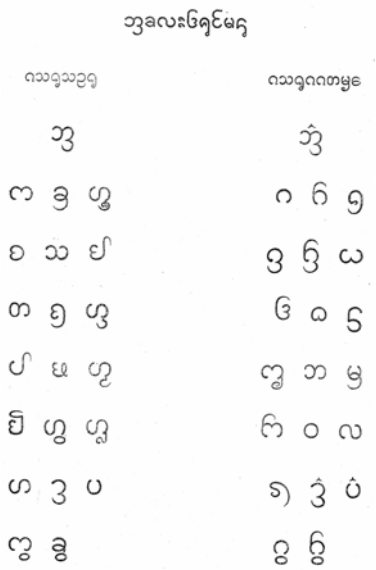

Figure 5. Tai Lue consonant chart: high and low class consonants listed separately (LA1, p. 7)

This same book, which is a literacy primer used for Tai Lue students in schools in the Xishuangbanna region, teaches the consonants with high- and low-class consonants interleaved, however. Unfortunately, it is not consistent in how it does this: the first eight consonants consider tone class after manner of articulation:

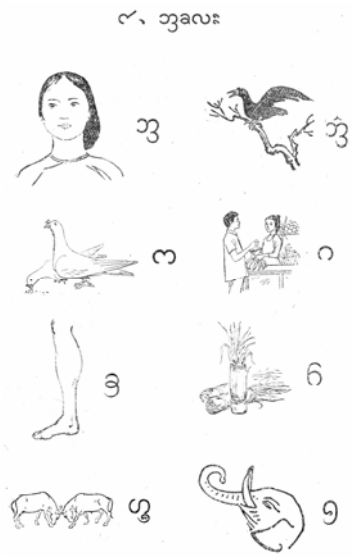

Figure 6. Tai Lue consonants in LA1: glottal and velar consonants (LA1, p. 1)

The ordering for the remaining consonants, however, considers tone class after point of articulation but before manner of articulation:

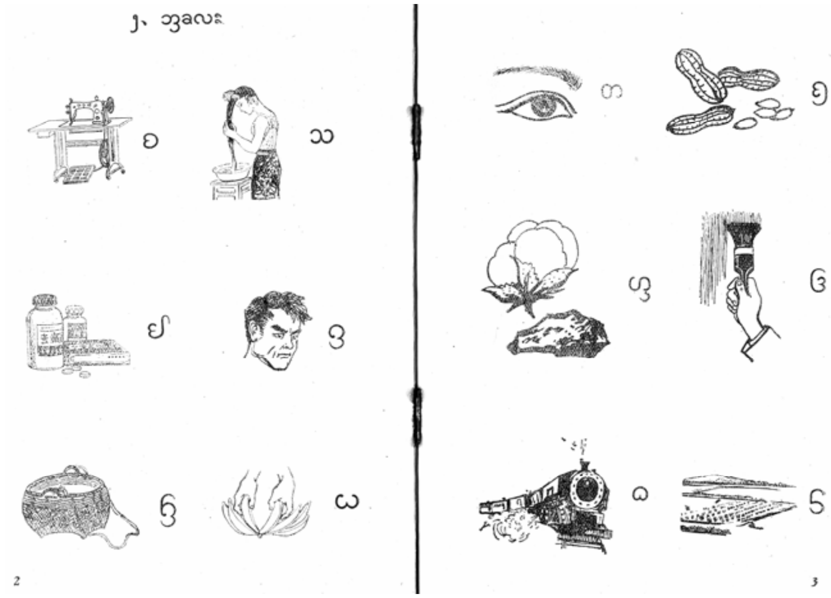

Figure 7. Tai Lue consonants in LA1: palatal and alveolar consonants (pp. 2–3)

It seems clear, though, that consonants are *not* taught using an order that considers tone class first (listing all high-class consonants before any low-class consonants). Also, in no case that I am aware of do the sort orders of other Tai scripts consider tone class before other consonant properties, point of articulation in particular.

Thus, it cannot be said for certain that the ordering of consonants in N2634 is the established conventional order. On the other hand, the ordering in N2242R, which considers tone class before point or manner of articulation, is not clearly attested in Tai Lue sources I have seen, and is unusual for Tai scripts.

2. **Ordering of diphthongs.** As noted in N2660, I consider the internal ordering of diphthongs in N2634 to be wrong, and suggest that IY should follow all other diphthongs. I concur, therefore, with the Chinese comments in this regard.

3. **Ordering of syllable-final consonants in relation to vowels.** N2242R lists syllable-final consonant characters in the middle of the vowel characters, after “ᦺ” /ai/ and before “ᦻ” /aːi/. In contrast, N2634 lists the syllable-final consonant characters before any vowels.

   The ordering of N2242R in this respect is unusual, and not supported in any source materials I have seen. For a given initial consonant C, it is unclear from sources if there are established conventions as to whether a syllable written in the form C + V should sort before or after a syllable in the form C + Cf; for example, ᦎᦰ < ᦎᦺ < ᦎᦻ < ᦎᧂ < ᦎᧃ versus ᦎᧂ < ᦎᧃ < ᦎᦰ < ᦎᦺ < ᦎᦻ. Syllable rhyme charts are not clear or consistent in this regard: the chart in Figure 8 indicates that C + V syllables precede C + Cf or C + V + Cf syllables. The chart in Figure 9 lists syllable-final consonants after vowels, but does not make clear what the ordering of syllable rhymes should be. The chart in Figure 10 shows the same ordering of closed rhymes as the chart in Figure 8, but provides no information regarding open syllables in relation to closed syllables.

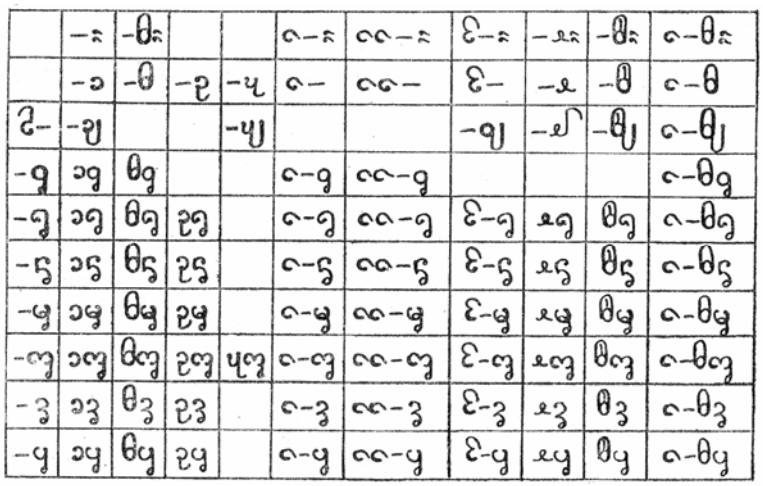

Figure 8. Tai Lue syllable rhyme chart: (LA1, p. 36)

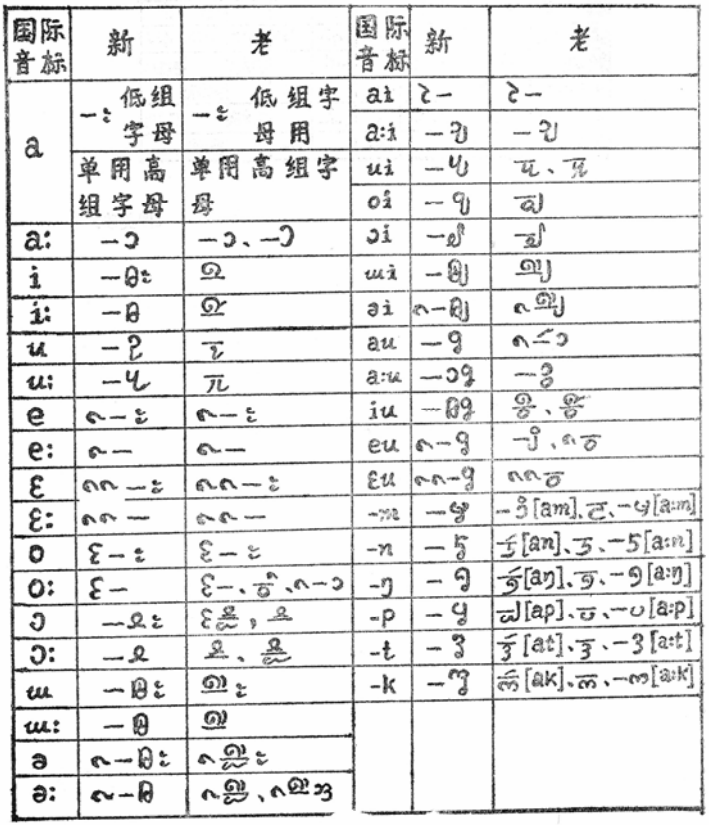

Figure 9. Tai Lue syllable rhyme chart: (SCEM, p. 73)

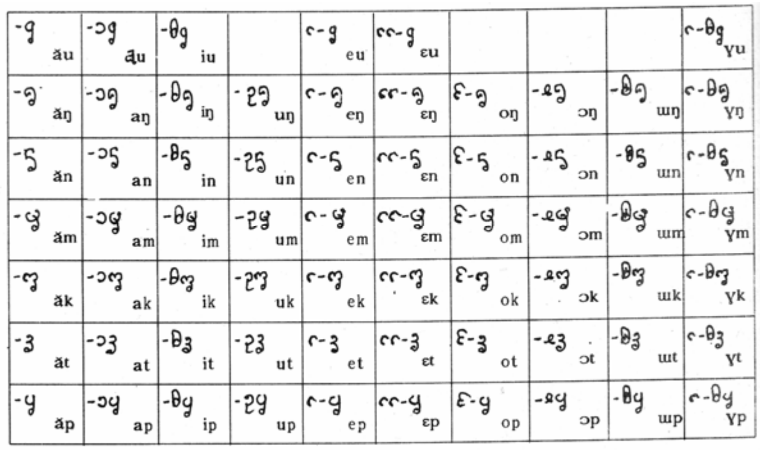

Figure 10. Tai Lue syllable rhyme chart: (PSCC, p. 166)

It is clear, however, that C + Cf syllables precede C + V + Cf syllables. It also seems clear that the order suggested by N2242R with C + Cf syllables after some C + V syllables and before others, e.g., ᦎᦰ < ᦎᦲ < ᦎᧂ < ᦎᧄ < ᦎᦳ, is not generally used.

Thus, while there is some uncertainty regarding conventional ordering of rhymes, there is some evidence to support the placement of syllable-final consonants before vowels, as in N2634, though the evidence equally supports placement of syllable-final consonants after all vowels; but I have not seen any evidence to support placement of syllable-final consonants after some vowels but before others, as in N2242R.

In summary, there is variation among source materials regarding collation order. Moreover, Tai Lue sorting involves considerations that go beyond mere code point ordering of characters. Nevertheless, with the exception of the internal order of diphthongs, I consider there to be good basis for adopting the ordering of characters in N2634.

References

[SCEM] 中国社会科学院民族研究所，国家民族事务委员会文化宣传司(Zhongguo shehui kexueyuan minzu yianjiusuo, guojia minzu shiwu weiyuanhui wenhua xuanchuansi = Chinese Academy of Social Sciences, Institute for Nationality Studies; Department of Cultural Propagation, State Ethnic Affairs Commission). 1991. 中国少数民族文字 (*Zhongguo shaoshu minzu wenzi = Scripts of Chinese Ethnic Minorities*). 北京 (Beijing): 中国藏学出版社 (Zhongguo zangxue chubanshe = China Tibetan Studies Publishing House).

[PSCC] 戴庆厦 许寿椿 高喜奎 主编 (Dai Qingsha, Xu Shouchuan, Gao Xikui, editors-in-chief). 1991. 中国各民族文字与电脑信息处理 (*Zhongguo ge minzu wenzi yu diannao xinxi chuli = Processing the scripts of China on the computer*). 北京 (Beijing): 中央民族学院出版社 (Zhongyang minzu xueyuan chubanshe = Central Nationalities Institute Publishing House).

ᦺᦎᦵᦈᦲᧃᦵᦃᦃᦵᦉᦲᧂ (搜集整理者 = Dao Jinxiang). 1994. ᦙᦸᦔᦱᦔᦃᦃᦐᦵᦗᦳᦀᦔᦃᦟᦲᦙ (古傣语词语译翻 = *Dictionary of ancient Dai words and expressions*). 昆明 (Kunming): ᦜᦲᧂᦵᧂᦱᧂᦵᦔᦲᧂᦵᦉᦌᦵᦉᦌᦵᦙᦱᧂᦃᦱᦲᦵ (云南民族出版社出版 = Yunnan Nationalities Press).

[LA1] 西双版纳州教育局编 (Xishuangbanna Prefectural Education Bureau). 1980. ᦀᦔᦡᦱᦙᦟᦲᦙᧃᦒ, ᦙᦳᦏᦵ ᦜᦲ ᦵ (语 文, 第一册 = Language arts, book 1). (西双版纳傣文小学课本 = Xishuangbanna Dai primary school textbook.) 昆明 (Kunming):云南民族出版社出版 (Yunnan Nationalities Press).

L2/99-243 Constable, Peter. 1999. “Proposal for encoding New Tai Lue script in Unicode/ISO-IEC 10646.”

N2242R China. 2000. “Proposal for encoding Xishuang Banna Dai script on BMP of ISO/IEC 10646.”

N2634 Eversion, Michael. 2003. “Proposal to encode the Tai Lue script in the BMP of the UCS.”