ISO/IEC JTC1/SC2/WG2 N4xxx
UTC L2/14-090
2014-04-23

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation

Doc Type: **Working Group Document**
Title: **Proposal to Change the Encoding Model of New Tai Lue**
Source: **Martin Hosken**
Status: **Individual contribution**
Action: **For consideration by UTC**
Date: **2014-04-23**

Executive Summary This proposal is to change the encoding model for New Tai Lue from being logically ordered with prevowels stored after the initial consonant, to visually ordered with the prevowels being stored before the initial consonant, as per Thai, Lao and Tai Viet. The easiest way to mark this change is to recategorise the characters U+19B0 .. U+19C0, U+19C8 ..U+19C9 from a general category of Mc, to Lo. Collation contractions also need to be added.

Introduction The New Tai Lue script has been encoding in Unicode for nearly 10 years now (v4.1 2005), but it has seen very little use among the primary user community within China. Indeed, the only use from Chinese users in Xisuangbanna, has been either using legacy encoded fonts or using a Unicode font where the behaviour is such that the reordering characters are stored in visual order. For example http://www.dw12.com is a news site using fonts with such an encoding.

Why is this an issue? Surely there are examples of other scripts where people have played fast and loose with Unicode, producing fonts that don't conform to the Unicode standard while using its codepoints[1], so why do we need to do anything about New Tai Lue? The reason for concern is that such user communities have inadvertently produced a fork of Unicode that, unlike for other scripts, could well stick. We are in a position now where users of New Tai Lue have to decide whether to stick with the Unicode Standard as specified or follow the community in its modified encoding model. To understand the depth of this issue, we need to understand how we got to where we are now.

Causes For users of the New Tai Lue script, a primary usability concern is that they can interact with their text as if it were visually ordered. Thus prevowels must be typable before the consonant they visually precede. Other concerns are secondary: treating prevowels as individual characters; accurate sorting. Another concern for users is the ease of making their own fonts. In terms of how readers consider the script, of course prevowels should be stored before the consonants. That's where they go. There are no diacritics, no ordering, text is written, typed and stored in simple linear order just like Latin scrip; just as in English, we don't reorder our vowels to put them together in a syllabel or sied by sied in a word. In this respect they follow Thai, Lao and Tai Viet. The fact that the script works linguistically either way (logically or visually ordered) only helps to muddy the waters. With respect to sorting, New Tai Lue follows Burmese and old Lao in sorting by Initial consonant, Medial, Final consonant, Vowel and Tone. Thus any sorting that doesn't handle the Final consonant emphasis, is a compromise and poor justification for a logical ordering.

The seeds of the problem we now face were evident even before New Tai Lue was added to the standard. In a unicore email of 11/Oct/2002, Peter Constable stated:

> The one issue on which SIL has a strong objection to the current proposal is the matter of handling re-ordrant vowel marks. The latest proposal (the URL I had for this no longer seems to work) handles these like combining marks, comparable to Devanagari i. We feel

1 Zawgyi fonts in Burmese, http://www.tai12.com for Lanna are such examples

very strongly that this script should use the Thai / Lao approach in which such vowels are Lo. Our reasoning is that the script revision that led from Lanna to New Tai Lue was clearly intended to result in a simple script. Making these to be combining marks that require reordering is the only thing that would impede relatively easy rendering implementations, and would delay any widespread implementation by many years, in our opinion.

In effect, data entry concerns have been the primary motivator in the refusal to use logical ordering by the user community. Consideration of these issues impacts the options open to resolving the issue of the parting of the ways in terms of Unicode usage

Solution 1 A natural reaction to such news might well be to say: We should get them to use Unicode as it was intended to be used. This is certainly a possibility, but are we willing to pay the cost? Microsoft, due to their near monopoly in the area, would bear the brunt of this. They would need to produce a keyboarding system that would allow data entry in the desired user order. An added, but insufficient on its own, incentive would also be to make the sequence U+0020 followed by a prevowel be illegal in a visual way. These would also all have to be back ported to Windows 7 and Office 2003! Would either of this approaches work on its own? Fixing data entry would provide an adequate path forward for users. Merely blocking rendering would have the effect of pushing users into legacy encodings rather than encouraging them to solve the data entry problem themselves. Assuming that Microsoft are not willing to address the keyboarding issue (having refused to for many years), I have to assume that this solution is not an option.

Solution 2 The natural Unicode solution is to deprecate the offending characters and introduce new ones:

U+19B5	→	U+19CA
U+19B6	→	U+19CB
U+19B7	→	U+19CC
U+19BA	→	U+19CD

This is an acceptable solution although it would involve everyone in a textual transition that could last a while.

Solution 3 A radical solution is to accept in its entirety what is happening in China. We simply declare that the prevowels are to be stored in visual order. Very little needs to happen with regard to changing anything in the core Unicode standard. It would help if all the characters with a general category of Mc were changed to Lo. This would reflect the attitude of users that every character is just that, a letter. There is no combining behaviour. The cost here is that there are implementations in existence that use the logical order and they would have to change. If that cost is acceptable, then this solution causes the least disruption to users.

In addition, this change requires no additions or removals from ISO10646. But since it does involve a significant model change, it is suggested that this change also come before ISO/IEC JTC1/SC2/WG2.

Conclusion There may be other approaches that can be thought of, but of those stated here, the best solution is to simply restate New Tai Lue as a visually ordered script. A summary is given in the executive summary at the start of this document. The full details are:

1. Change the General Category of characters U+19B0 .. U+19C0, U+19C8 ..U+19C9 from Mc to Lo.

2. Collation. Tai Lue is typically sorted in the priority order of: Initial cluster, final consonant, vowel, tone. Supporting this in the DUCET would result in too many contractions. Thus a compromise is proposed for the DUCET which would give Tai Lue a fallback sorting to that akin to Thai. For this the following sequence multiplication would be required:

([U+19B5, U+19B6, U+19B7, U+19BA])([U+1980-U+19AB]) = \2\1

adding 172 entries to the DUCET.

For example, consider the first result of the multiplication: U+19B5 U+1980. We want this to sort as if it were U+1980 U+19B5. In ICU terms this would be done by:

&\u1980=\u1980\u19b5/\u19b5.

A full tailoring of 5050 contractions for Tai Lue can be provided.