Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Comments on proposed Vietnamese Reading Marks**
Source: **Andrew West, John Knightley, Eiso Chan**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2018-08-30**

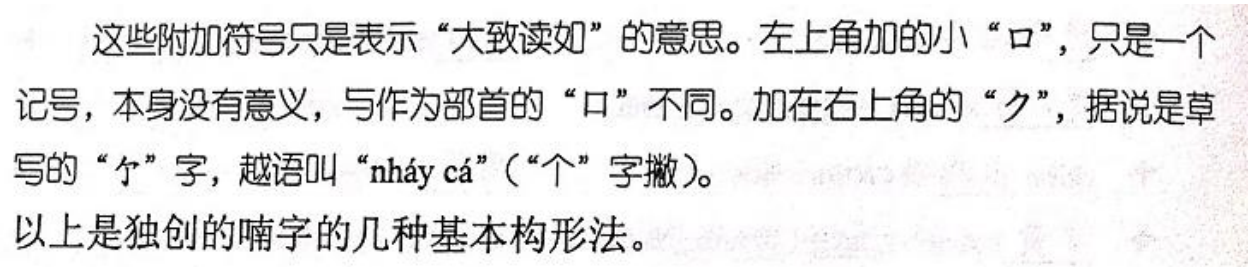
1. Discussion

This document is a response to the *Proposal to Encode Two Vietnamese Alternate Reading Marks* by Lee Collins (WG2 N4915, L2/17-373R). Collins' document proposes encoding two diacritical marks for use with CJK unified ideographs used in Vietnamese Chữ Nôm script: *cá* 𠂆 and *nháy* 〈. We have reviewed the proposal, and consider that it is problematic, and should not be accepted in its current form.

A. Reading Marks

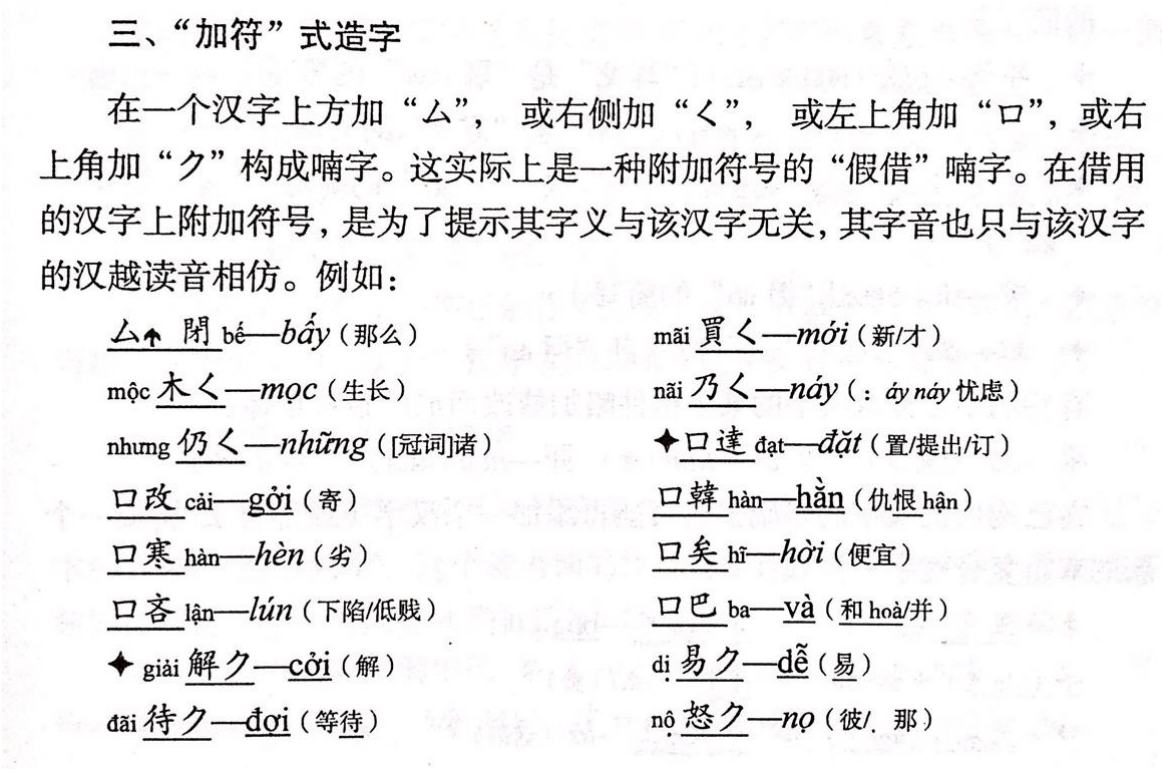
We accept that these two marks may be used as reading marks, indicating that the character is to be read like the base character, but they are certainly not the only reading marks used for Vietnamese. Reading marks are also used in China for writing Vietnamese, Zhuang and other minority languages using Han ideographs. Other Vietnamese reading marks noted by Chinese scholars include 𠂇 on the right of a base character and 𠘨 on top of a base character (see Fig. 1 and Fig. 2). The component 乙 in U+2B86F 𫚭 and other V-source CJK unified ideographs is also a kind of reading mark, as is the component 刂 (and variants) on the left side of a character.

Fig. 1: Discussion of reading marks in Luó and Xíng 2013



这些附加符号只是表示“大致读如”的意思。左上角加的小“口”，只是一个记号，本身没有意义，与作为部首的“口”不同。加在右上角的“𠂇”，据说是草写的“个”字，越语叫“nháy cá”（“个”字撇）。
以上是独创的喃字的几种基本构形法。

Fig. 2: Examples of reading marks used to create new characters in Luó and Xíng 2013



三、“加符”式造字

在一个汉字上方加“厶”，或右侧加“く”，或左上角加“口”，或右上角加“ク”构成喃字。这实际上是一种附加符号的“假借”喃字。在借用的汉字上附加符号，是为了提示其字义与该汉字无关，其字音也只与该汉字的汉越读音相仿。例如：

厶↑ 閉 bế—*bấy*（那么）
mãi 買く—*mới*（新/才）
mộc 木く—*mọc*（生长）
nãi 乃く—*náy*（：*áy náy* 忧虑）
nhưng 仍く—*nhưng*（[冠词]诸）
✦口達 đạt—*đặt*（置/提出/订）
口改 cải—*gởi*（寄）
口韓 hàn—*hằn*（仇恨 hận）
口寒 hàn—*hèn*（劣）
口矣 hĩ—*hời*（便宜）
口吝 lận—*lún*（下陷/低贱）
口巴 ba—*và*（和 hoà/并）
✦ giải 解ク—*cởi*（解）
dị 易ク—*dễ*（易）
đãi 待ク—*đợi*（等待）
nộ 怒ク—*nọ*（彼/，那）

In particular, a small square mark placed at the top left of an ideograph is also a reading mark, and there are dozens of CJK unified ideographs with a single Vietnamese source that use this square reading mark (see Fig. 3 for a few examples). In these cases the square reading mark is treated as the ‘mouth’ 口 radical, although technically it is not a radical. The square reading mark has a long history, and in the 12th-century Tangut-Chinese handbook *Pearl in the Palm* (番漢合時掌中珠), a square reading mark at the top left is used in many Han characters used for transcribing Tangut, although in modern scholarship it is treated as a mouth radical. It is also used in Buddhist usage, as well as for Cantonese (e.g. in the characters 嗰, 哋, 唅), and for Zhuang, Bai and Miao characters.

Fig. 3: Excerpt from Code Chart for CJK Unified Ideographs Extension B

20F3F 口 30.11	噁 UCS2003	噁 V2-7145
20F40 口 30.11	嘔 UCS2003	嘔 V2-714F
20F41 口 30.11	噓 UCS2003	噓 V2-7159
20F42 口 30.11	噝 UCS2003	噝 V3-3055
20F43 口 30.11	噢 UCS2003	噢 V0-326B
20F44 口 30.11	噺 UCS2003	噺 V0-3326
20F45 口 30.11	嚏 UCS2003	嚏 V2-7146

The component 〈 is also treated as a radical by some authors, for example in the 2016 handbook of Nom characters used by the Jing people in China (see Fig. 4).

Fig. 4: 〈 Radical in Hé Sīyuán 2016

〈部

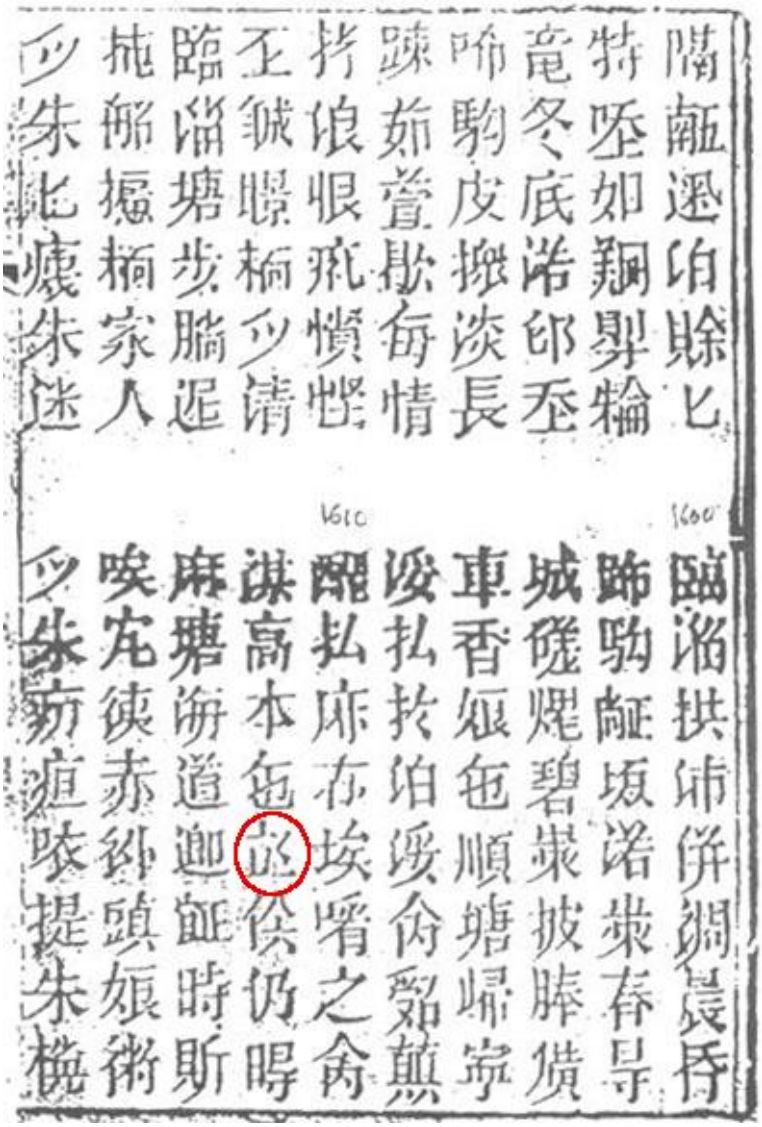
笔画数	喃字	国际音标	拉丁越南文	字义
5	尼〈 ①	nai^{2}	này	此；该；这个，这些；就在这里
5	尼〈 ②	noi^{1}	nôi	衷，内心，
5	尼〈 ③	nəəi^{1}	nơi	地方，……之处
5	外〈	ŋaai^{2}	ngài	外
7	来〈	joi^{2}	rồi	了；结束
8	练〈	jɛn^{5}	rén	蹑手蹑脚
8	臥〈	kɔ^{5}	có	有
12	斲〈	kat^{7}	cắt	同“刧”“刹”“劫”：割
12	運〈	vən^{2}	vần	运

B. Diacritical Marks or Character Components?

We do not agree with the assertion that these two marks are diacritical marks. Diacritical marks should normally be positioned on top of, below, or to the side of a base character, without affecting the graphic form of the base character. However, in printed sources that we have examined, these two marks are not positioned to the side of a base ideograph, but are positioned within the ideographic box, causing the base character to be compressed laterally to make space for the mark. For this reason, many scholars consider these reading marks to be integral components of characters they occur in, and building blocks for the construction of new characters.

The fact that 〈 is not a diacritical mark can be seen in the 1872 edition of *Truyện Kiều Bản* shown Fig. 5, where the reading mark 〈 occupies the right half of the ideographic square of the character 立〈, and the base character 立 is compressed accordingly. In the Nom Na Tong font provided by the Nom Foundation, this character (U+F0865) also shows lateral compression of 立, although the reading mark is smaller and the compression of the base character less than in the source edition: 立〈 (compare 立 in the same font). This means that a combining reading mark cannot be simply overlaid on an existing ideograph to produce the correct glyph form, but a font would need substitute the sequence of a CJK unified ideograph followed by a combining reading mark with a precomposed glyph in which the two components are harmoniously positioned within the same ideographic box used for any other CJK unified ideograph.

Fig. 5: Folio 42a of the 1872 edition of *Truyện Kiều Bản*



Source: http://nomfoundation.org/nom-project/tale-of-kieu/tale-of-kieu-version-1872

Another example of the lateral compression of the base character to accommodate the reading mark 〈 can be seen in the example from the 2009 dictionary *Từ Điển Chữ Nôm Trích Dẫn* shown in Fig. 6. Here there are entries for both 意 and ⿰意〈, and the 意 component of ⿰意〈 is noticeably thinner than 意 as a separate character. Clearly, 〈 is not being treated as a diacritical mark overlaying the character 意, but as a component in the construction of a new character.

Note that the character with 𠂉 corresponding to ⿰意〈 was encoded in Ext. C as U+2AB2B 𪜫 (see Fig. 7). Likewise, ⿰立〈 is also written using the 𠂉 mark instead of the 〈 mark, and was encoded in Ext. C as U+2B05E 𫁞. In the case of U+2B05E, not only is the 立 component laterally compressed, but it has the special left-side radical form of the character 立. If 𠂉 was a diacritical mark the base character would not take the left-side radical form.

2B05E
立 117.3
𫁞
V4-4F7E

Fig. 6: ***Từ Điển Chữ Nôm Trích Dẫn*** **(Dictionary of Nôm characters with excerpts) p.19**

TĐCNTD 字典節喃摘引 ấy

意 **ấy**

Bộ: 心, **Nét:** 61.10-14

❶ **Âm:** 意 (ý), く : **dấu nháy.** ✿ 小姐吏咿祕娘、曷慍掀曲断腸意之。Tiểu thư lại thét lấy nàng: "Cuộc vui gảy khúc đoạn trường *ấy* chi?" (Kiều KOM, c. 1859-1860).

意 **ấy**

(ý, e)

Bộ: 心, **Nét:** 61.9-13, **U+610F**

☞ 意: ý (trang 1494).

Source: https://www.chunom.org/pages/tdcntd/#19

Fig. 7: ***Tự Điển Chữ Nôm Dẫn Giải*** **(Dictionary of Nôm Characters with Quotations and Annotations) p.58**

ấy
𢧐
V+636F0

#{C2>G2: ý 意>𢧐}. ◎ (*như trên*). 台茹𢧐氣象習俗体咟拯同 Hai nhà ấy khí tượng tập tục thảy cả chẳng đồng (Truyền kỳ, I, Khoái Châu, 15b) ○ 埃可皮意𢧐付默觧春撽撕 Ai khả vừa ý ấy, phó mặc gió xuân đun đẩy (Truyền kỳ, III, Thúy Tiêu, 44b) ○ 才能𢧐𢆶荣華𢧐 唆嚫賒斯乙參稽 Tài năng ấy sánh vinh hoa ấy. Lừng lẫy xa gần ắt chỉn ghê (Bách vịnh, 39b) ◎ *Tiếng tỏ ý chuyển tiếp, nối kết.* 𢧐𢧐玄機揜掣嗐 Ấy ấy huyền cơ chẳng xiết khen (Bách vịnh, 37a) ○ 𢧐些祕意耒咟干烝眉 Ấy ta lấy ý vua, cả can chưng mày (Thi kinh B, IV, 42b).

B. Existing CJK Unified Ideographs with Reading Marks

There are fifty-seven existing V-source CJK unified ideographs with the reading mark 个, 𠂉 or 𠂊 on the right side: 31 in Ext. C (2A76A, 2A771, 2A780, 2A7F3, 2A809, 2A849, 2A84D, 2A8F1, 2A932, 2A938, 2A97E, 2AA6C, 2AAD3, 2AAF6, 2AB2B, 2ABAF, 2ABCA, 2AC07, 2AC30, 2AC55, 2ACA3, 2AD8F, 2AE04, 2AF86, 2AFF8, 2B05E, 2B1A1, 2B27C, 2B2F6, 2B391, 2B39A); and 26 in Ext. E (2B850, 2B896, 2B9E3, 2B9E6, 2BD2C, 2BD9C, 2C086, 2C0A8, 2C192, 2C2C6, 2C2EB, 2C2F1, 2C323, 2C3BB, 2C3BD, 2C438, 2C4D7, 2C573, 2C57A, 2C5A9, 2C7AE, 2C89E, 2CA3B, 2CA84, 2CABD, 2CB03). In all cases the reading mark is treated as a component of the character, taking up space in the ideographic square, rather than as a diacritical mark appended to the side of the ideographic square. Encoding the reading mark 𠂉 as a combining mark means that there would be two different ways of representing the sixty already-encoded characters, and the two representations would not be canonically equivalent.

2. Conclusion

There is no fundamental difference between the reading marks 𠂉 and 〈 and other reading marks such as the square mark positioned at the top left, and in the same way that the square reading mark is treated as a character component for encoding purposes, 𠂉 and 〈 should be treated as character components rather than as diacritical marks.

There is no advantage for font designers to encoding 𠂉 and 〈 as combining diacritical marks because they are not diacritical marks but integral character components, and so fonts would still need to provide precomposed glyphs for each supported sequence of base ideograph and combining mark.

As fifty-seven V-source CJK unified ideographs with the reading mark 个, 𠂉 or 𠂊 have already been encoded, encoding a new combining mark for this reading mark would introduce duplicate representations which are not canonically equivalent. This would be problematic for searching and text processing, and would cause confusion for end users.

Whilst no V source CJK ideographs that use 〈 as a reading mark have been encoded to date, encoding 〈 as a combining mark would still be problematic. We would like to see analysis of how many characters incorporate the 〈 reading mark. If it is not a very large open-ended number, then encoding the characters separately may be preferable to encoding 〈 as a combining mark. We would also like to see further discussion of other reading marks such as 𠂇 and 𠂆, in order to better understand the scope and usage of reading marks, and whether other combining marks may be required if combining marks are accepted as the solution.

Encoding combining marks as a method for extending the encoding of CJKV unified ideographs is a major innovation, and is a step that should not be taken lightly. We recommend that more extensive review of this proposal be carried out by experts on CJKV ideographs, and that the proposal is discussed at IRG Meeting 51 in October 2018. We think that there should be consensus among IRG members before any combing marks for use with CJK unified ideographs are accepted for encoding by UTC or WG2.

3. Bibliography

Hé Sīyuán 何思源. 2016. *Zhōngguó Jīngzú Nánzì Hànzì Duìzhào Shǒucè* 中国京族喃字汉字对照手册. Beijing: Minzu Chubanshe. ISBN 978-7-105-14624-6

Luó Qǐhuá 罗启华 and Xíng Fúyì 邢福义. 2013. *Yǔyán de qīnqíng — Yuènányǔ Hànyuán Chéngfēn Tànxī* 语言的亲情——越南语汉源成分探析. Wuhan: Huazhong Shifan Daxue Chubanshe. ISBN 978-7-5622-5647-2

Nguyễn Quang Hồng. 2014. *Tự Điển Chữ Nôm Dẫn Giải*. Hà Nội: Nhà xuất bản Khoa xọc Xã hội and Bắc Carolina: Hội Bảo tồn Di sản Chữ Nôm (North Carolina: Vietnamese Nôm Preservation Foundation). ISBN 978604902283