L2/18-318
WG2 N5026

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

Doc Type:	**Working Group Document**
Title:	**Response to L2/18-281 (Comments on proposed Vietnamese Reading Marks)**
Source:	**Lee Collins, Vietnamese Nôm Preservation Foundation** **Ngô Thanh Nhàn, Temple University**
Status:	**Individual Contribution**
Action:	**For consideration by JTC1/SC2/WG2 and UTC**
Date:	**2018-10-20**

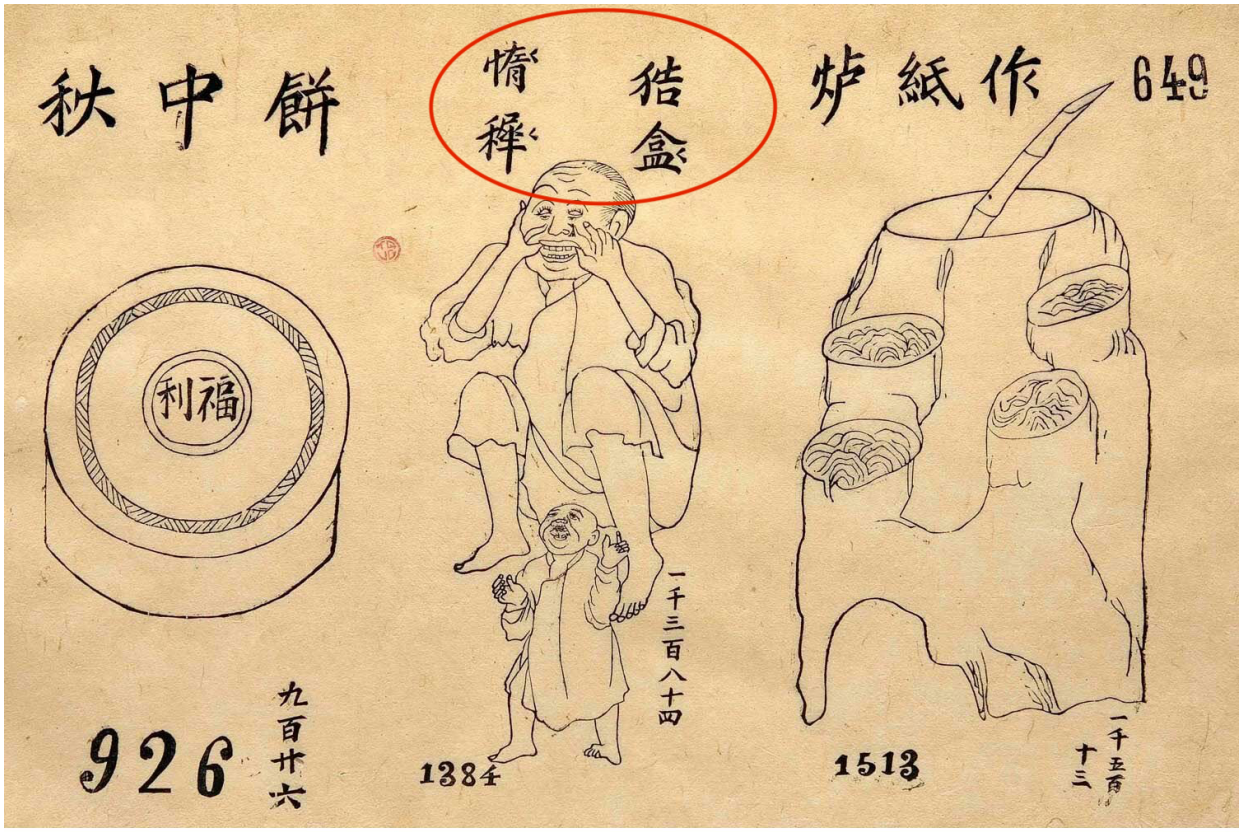
Introduction

In **L2/18-281,** Andrew West, John Knightley, and Eiso Chan, have questioned our proposal to include 2 Vietnamese reading marks in Unicode, considering our proposal "problematic". We argue that their paper reveals a fundamental misunderstanding of the nature of the Unicode character-glyph model and of how the Vietnamese view and use the two diacritical marks. In this paper, we will attempt to address some of their misunderstandings.

We first note that they failed to note that our paper was the joint contribution of two authors, Lee Collins, and Dr. Ngô Thanh Nhàn of Temple University.

Diacritical Marks vs. Character Components

The crux of the argument in **L2/18-281** is that these are not marks because some renditions are designed to fit within the imaginary box that bounds a CJKV Ideographic character. The authors provide a few examples, but conveniently ignore counter examples such as those shown in the figure below where *nháy* marks in a single text can be seen drawn outside and within the bounding box. The correct conclusion to draw from this is that the position of the marks is not critical to their nature as diacritics.



Thus, **L2/18-281** confuses an artifact of design determined by a particular font and written medium (e.g. Song font on woodblock) with the semantics (meaning, intention, application) of the diacritics. We and the Vietnamese scholars cited in the bibliography argue that in order to determine whether an element is a combining mark, one should avoid using neatly printed or carved ideograms from modern documents, because they have already been regularized. Rather, one should go to the hand-written materials, and study how Vietnamese writers use the marks and the conclusions drawn about their use by Vietnamese scholars.

The Unicode character-glyph model in fact allows great flexibility in the realization of the rendered form of an encoded character sequence. This is frequently illustrated by models from the Devanagari script. For example, the matra ु *u* in Devanagari is encoded as a separate combining mark. However, when it is rendered under the letter र *ra* it is normally tucked within the bounding box of the *ra* to achieve रु. The two characters are encoded separately and rendered as one shape dynamically. A more extreme example is the character cluster *kṣa.* In Unicode it is encoded as 3 separate characters क ् ष, but rendered as the single element क्ष.

Similarly, the fact that a font designer places the *nháy* diacritic within the bounding box of the modified glyph is purely a typographical or aesthetic consideration. It does not change the underlying semantic or relationship of the two elements.

In sum, the shape based arguments listed in **L2/18-281** against the reading marks fall apart when seen merely as an application of the Unicode character-glyph model.

The nature of the reading marks

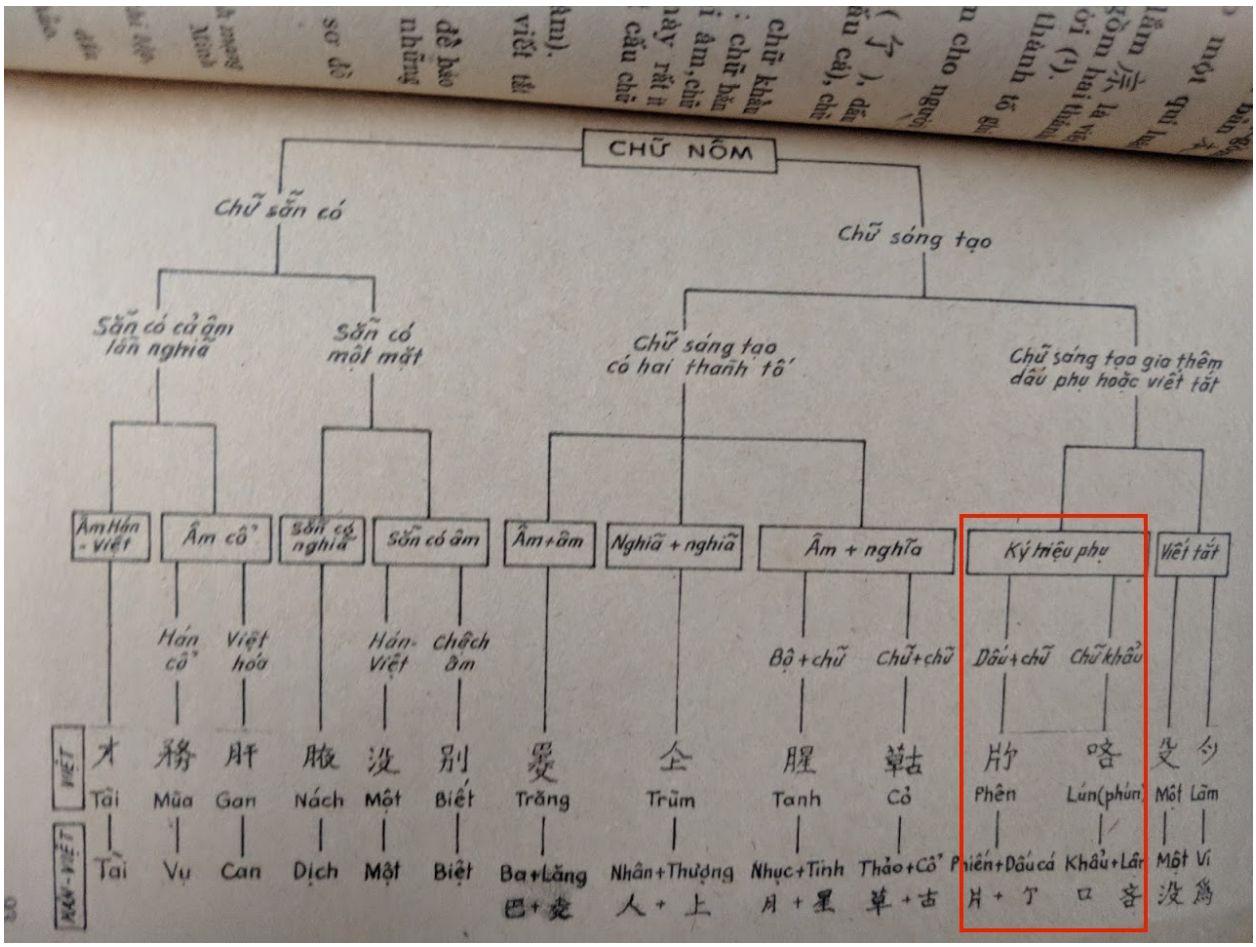
Following the the majority of Vietnamese scholars, we argue that how Vietnamese writers understand and use these marks is more important to determing their encoding than the final shape. Standard views can be found in the works listed below: Đào Duy Anh 1975, Lê Văn Quán 1981, and Vũ Văn Kính 1994.

These authors argue for a model in which Vietnamese distinguish between ideograms proper (*chữ* 𡨸), radicals and other components of characters (*bộ* 部), and diacritical marks (*dấu*) applied to those characters.

Vietnamese scholars choose the word *dấu* because it is the generic word for *diacritical mark*. The term *dấu* is also used to denote the tone marks in the Vietnamese *quốc ngữ* script. For example, *dấu hỏi* is what Vietnamese call the tone mark encoded in Unicode as U+0309 COMBINING HOOK ABOVE. *Dấu*, then is a mark that can be freely attached to a character, not a component.

There are several reasons for singling out the usage of *Dấu*. One is that they indicate that a normal Hán character is being used with a native Vietnamese sense. Secondly, this usage is optional, so we can find instances Hán characters marked and unmarked by *Dấu* in a single text. (Lê Văn Quán p. 92). Thus, *cá* and *nháy* are explicitly not called "*Radical Cá" or* "*Radical Nháy"*. They are called *Dấu* because understood to be add-ons, separate from the base character.

For an overview in schematic form, see the Figure below from p. 93 of the study by Lê Văn Quán.



Prof. Quán distinguishes the category of *ký hiệu phụ* (annotational marks) as a separate method of chữ Nôm formation. *ký hiệu phụ* themselves are divided into *Dấu* (diacritical marks), currently under consideration, and *chữ khẩu* (the character 口). Prof. Quán notes that the Hán character *khẩu* can be used as an annotational mark.[1]

The Reading Marks are not an innovation

L2/18-281 states that "Encoding combining marks as a method for extending the encoding of CJKV unified ideographs is a major innovation.**"**

This is wrong on three counts:

1. As we have argued previously and above, this proposal is not a method to extend the encoding of CJKV unified ideographs. Rather, it is a method to mark encoded ideographs as having a demotic reading. Encoding them as combining marks is merely to achieve the correct editing behavior in applications. Also, since the marks can be applied to virtually any Hán character (see appendix one below for examples), they are encoded as marks for maximum flexibility, just as Unicode allows accent marks to be attached to any Latin script base letter.

[1] In fact, it might make sense to treat most cases of *khẩu* in *chữ Nôm* as reading marks. However, the weight of encoding practice in the various CJKV standards and Unicode has been against this.

2. Using combining marks for *cá* and *nháy* is not new. As noted previously, it is a method introduced over 25 years ago in TCVN 5773-1993, the first national standard to encode Nôm.

3. Unicode itself already has an established precedent for diacritical marks to specify attributes of CJKV ideographs with the encoding of the four IDEOGRAPHIC tone marks (U+302A ... U+302D) as combining marks. Just as the tone marks do not encode a new character, but make explicit the inherent tone, the Vietnamese reading marks merely clarify the usage of the existing ideograph.

Encoding and compatibility issues

Regarding the encoding of the mark *cá* in particular , **L2/18-281** argues that "encoding a new combining mark for this reading mark would introduce duplicate representations which are not canonically equivalent."

This is not new information. We previously noted the existence of duplicate representations in our original proposal. While it is unfortunate that a relatively small number of pre-combined forms have been encoded, this is not a new problem in Unicode. Consider the many accented Latin script characters that have precomposed equivalents.

The precomposed forms with *cá* are a well-defined subset (see appendix two below). We do not see significant problems in excluding these from composition-decomposition as described in *Unicode Technical Report #15,* section 6. This is because in practice, most input of chữ Nôm characters happens in an input method, so that when users type the *cá* diacritic following a base character to create an an ambiguous sequence, the input method can replace the two characters with the pre-composed version if it exists. This is much easier to do than to try to get the system to handle combining sequences correctly. This is a minor inconvenience compared to the win of handling Nôm characters in a way that is intuitive and flexible.

The placement of the mark *cá*

In the *Comments on Public Review Issues L2/18-274*, John Knightly raises the issue of the placement of the *cá* diacritic. Citing an analysis presented in the Vietnamese review of IRG WG2017 https://hc.jsecs.org/irg/ws2017/app/index.php?id=05027, he notes that there is one case where the diacritic appears above the character. This is the character U+2B89A, 𫚭, which was encoded in Extension E.

In fact, in our database, which represents more than 30 years of research pertaining to the encoding of Hán-Nôm characters, there are only two cases where *cá* appears over the base character. One, U+2B89A, is already encoded. In the other known case, *rông*, the mark *cá* appears over the character 弄. The other ~160 known examples all place the *cá* to the right of the base character. We do not consider two anomalies as a reason to not encode the *cá* assuming its well established and productive position to the right of the base character. If at some time

large numbers of the mark *cá* are found to be in another position, we could consider encoding an additional varian. We do not consider this likely.

Bibliography

1. Đào Duy Anh. 1975. *Chữ Nôm: Nguồn gốc, cấu tạo, diễn biến* [Nôm script: Origin, formation, and development]. Social Sciences Publishing House.
2. Lê Văn Quán. 1981. *Nghiên cứu về chữ Nôm* [A study in Nôm ideograms]. Hanoi: Social Sciences Publishing House.
3. Vũ Văn Kính. 1994. *Bảng tra chữ Nôm miền Nam* [A glossary of Nôm ideograms used in south Vietnam]. Ho Chi Minh City: The Linguistic Society.

Appendix 1

370+ characters modified by Nháy

Appendix 2 Encoded characters with *Cá*

2A76A	𪝪
2A771	𪝱
2A780	𪞀
2A7F3	𪟳
2A809	𪠉
2A849	𪡉
2A84D	𪡍
2A8F1	𪣱
2A932	𪤲
2A938	𪤸
2A97E	𪥾
2AA6C	𪩬
2AAD3	𪫓
2AAF6	𪫶
2AB2B	𪬫
2ABAF	𪮯
2ABCA	𪯊
2AC07	𪰇
2AC30	𪰰
2AC55	𪱕
2ACA3	𪲣
2AD8F	𪶏
2AE04	𪸄
2AF86	𪾆
2AFF8	𪿸
2B05E	𫁞
2B1A1	𫆡
2B27C	𫉼
2B2F6	𫋶
2B391	𫎑
2B39A	𫎚
2B578	𫕸
2B850	𫡐
2B896	𫢖
2B9E3	𫧣
2B9E6	𫧦
2BD2C	𫴬
2BD9C	𫶜
2C086	𬂆
2C0A8	𬂨
2C192	𬆒
2C2C6	𬋆
2C2EB	𬋫
2C2F1	𬋱
2C323	𬌣
2C3BB	𬎻
2C3BD	𬎽
2C438	𬐸
2C445	𬑅
2C4D7	𬓗
2C573	𬕳
2C57A	𬕺
2C5A9	𬖩
2C7AE	𬞮
2C89E	𬢞
2CA3B	𬨻
2CA84	𬪄
2CABD	𬪽
2CB03	𬬃
2DA26	𭨦