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

Doc Type: **Working Group Document**
Title: **Proposal to Encode Two Vietnamese Alternate 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: **2017-11-6**

Introduction

This paper proposes the encoding of two diacritical marks used in the Chữ Nôm script to indicate alternate readings of Hán characters (*chữ Hán*). These marks (*dấu*) are collectively referred to in Vietnamese as *cá-nháy*, from the terms for the two most common marks, *cá* 个 (variant: 个) and nháy 〱 .

Chữ Nôm is composed roughly of two types of ideographic characters: characters borrowed directly from Chinese (chữ Hán) and characters created in Vietnam to represent native Vietnamese words (chữ Nôm)[1]. Figure 1 provides a convenient overview and further break down of this scheme[2].

The borrowed characters, *chữ mượn Hán*, were used to write the many words imported from a dialect of Middle Chinese, with original meanings and readings adapted into a Vietnamese standardization of that dialect (called *Hán-Việt* or Sino-Vietnamese, SV). They were also used without modification to represent native words and sounds.

For example:

日 sun, SV *nhật* > Vietnamese *nhặt* (pick up, gather, etc.)

---

[1] This includes words borrowed from Chinese during a very early stage, but which are considered native words, such as 鈽 *búa*, meaning *"axe",* derived from Chinese 斧
[2] Hông, 2014, p 17

白 white, SV *bạch* > Vietnamese *trắng* (white)

To overcome the inherent ambiguity in such a system, writers resorted to various schemes to distinguish Vietnamese from Sino-Vietnamese usage. Among these schemes, scholars distinguish two main categories[3]:

1. The application of diacritical marks to an original *Hán* character.

2. Composition of new characters containing phonetic and or semantic elements.

In the first scheme, diacritics are used to indicate that a borrowed character was being used as a native word, and that a native reading approximated by the Sino Vietnamese reading should be used[4]. In the second, some writers created new characters using phonetic elements to more directly specify the native reading.

To illustrate this difference using the two characters shown above:

1. the mark *nháy* was added to 日 to obtain 日ᐟ *nhặt*.

2. the phonetic element 壯 *tráng* was added to 白, creating 皨 *trắng*.

Similar to *nháy*, the mark *cá* 个 was also added to 日 to obtain 旷 *nhặt*, meaning "thick, dense".[5]

Additional examples of *cá* and *nháy* can be seen in figure 2.

Encoding history and usage

Cá and *nháy* were originally encoded in TCVN 5773-1993, the Vietnamese national standard for Chữ Nôm[6]. In defining them, TCVN 5773-1993 clearly distinguish them as "marks" not "characters". Specifically, TCVN 5773-1993 §5.3 and §5.4 of distinguish the encoding of "Non-combining" character from "Combined (sic) character":

---

[3] Hoang 2004, p. 1. Though as suggested here, sometimes the boundaries are not clear, for example the rare case where *cá* is used as a phonetic, e.g. 魛 *cá* (fish).

[4] Hồng 2013 pp 511-524, provides a detailed discussion of the different ways these marks were used

[5] Both examples in Hồng 2013, p. 1335.

[6] See figures 3 and 4 at the end of the proposal.

5.3 Non-combining character: member of the set of Nôm characters (except for 个 cá V+A171 and 丨 nháy V+A02F)

5.4 Combining character: a member of an identified subset of the coded character set of this standard, i.e 个 cá and 丨 nháy marks (V+A171 and V+A02F, respectively), intended for combination with non-combining character (sic)[7]

TCVN 5773-1993 calls out these two marks since they are the most productive. In the VNPF digital library alone we have identified close to 400 instances of characters combined with *nháy* and 160 combined with *cá*. We have also identified a small number of other marks. However, it is not yet clear that they are of sufficient productivity to warrant encoding. Further discussion may be found below.

While the non-combining unique characters of TCVN 5773-1993 were largely included in Unicode, for some reason, the *cá-nháy* were left out. It is not clear whether the omission was deliberate or merely an oversight, but it is an omission that needs to be remedied to facilitate the interchange of *chữ Nôm* information among user communities, such as the Facebook group 委班復生漢喃越南 (Ủy Ban Phục Sinh Hán Nôm Việt Nam).

Compatibility considerations.

Given the lack of the appropriate characters in Unicode, current practice to represent *nháy* is to type a LESS THAN SIGN, U+003C or the ideograph U+21FE8 following the base letter, while *cá* is represented either by U+4E2A 个 or the variant U+4E87 丁. None of these are satisfactory since they lack the properties necessary to support the desired clustering behavior. Mapping will be required to migrate from pseudo TCVN 5733-1993 to full Unicode, once the marks are encoded.

To our knowledge, no characters with glyphs containing *nháy* have been encoded in Unicode. However, a rather large number (64) of glyphs, such as U+2A76A composed with the mark 个 are already in Unicode. Compatibility decomposition mappings for these can be provided by the VNPF.

Behavior and positioning.

---

[7] TCVN 5773-1993 , p. 3

As with other combining marks, the *cá-nháy* should immediately follow the base character, in almost all cases a Hán Ideograph. In this respect, they are similar to the four IDEOGRAPHIC TONE MARKs encoded at U+302A-U+302D. Both marks are typically attached to the right of the base character, though variations of the *nháy* can be found above and below the center line of the base ideograph. They can also be integrated more closely into the shape of the base character as seen in the examples below. Font developers may wish to use positioning or ligature tables in their fonts to obtain a more precise appearance.

Shape and Variants

Nháy has two common forms. By far, the most common is that shown in the picture below, which can be found at http://www.chunom.org/pages/21FE8.


A variant form of *nháy*, typically found in texts from South Vietnam, uses the form seen below, applied to the character 庄:

The variants do not normally appear in a single text. If needed, they can be encoded using a variation sequence.

As noted above, *cá* has two forms. The source form is 个, but 个, a cursive rendition of 个, is much more common. Here is an example from the 1872 ed. of the Truyện Kiều, in which we see cá 个 added to 量 to mark the reading as *lưỡng*.
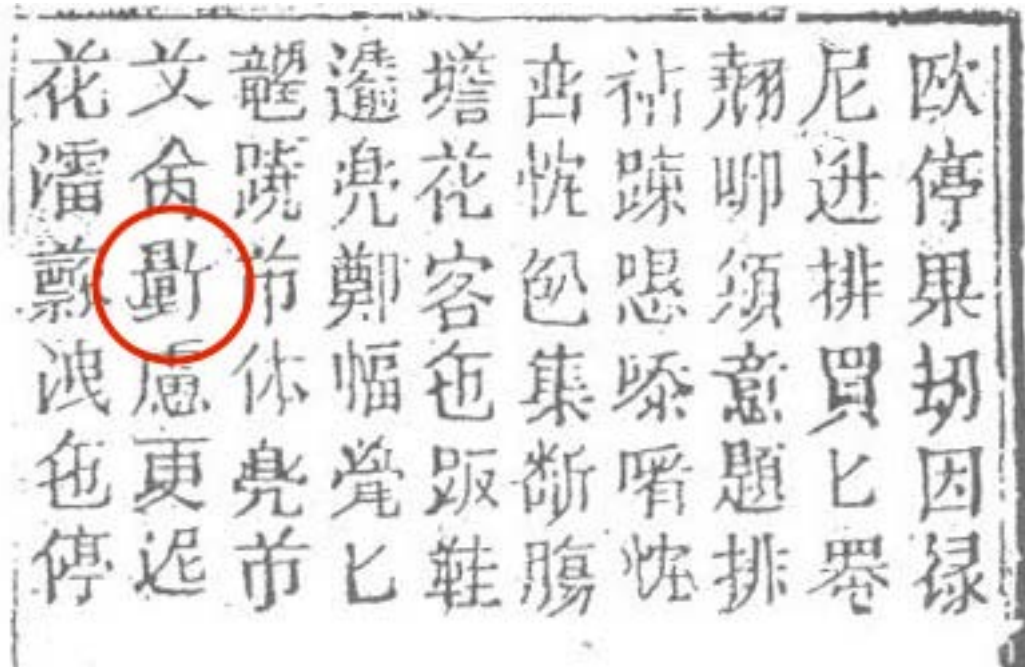
Our suggestion is to treat as 个 as primary, and 个 as a variation.

Other marks

Some scholars of Nôm argue for a larger set of marks. 4 are often cited[8],

1. the element 乁 in U+2B86E 逓.
2. the two dots on the either side of 力 in the character U+529E 办.
3. the element 刂 (and its variant 丷) in the character 順.
4. the two dots 丶 on the left side of the nôm Tày character 洎.

---

[8] See figure 5 below for examples

In most cases, these elements be shown to be abbreviations (*viết tắt*) of more complex characters — in these examples 霛, 辦, 鎖, and 咺, respectively. Also the other marks have not yet been proved to be as productive as *cá* and *nháy*. Therefore, they are not included in the current proposal. However, we do suggest leaving space for future additions.

Collation

Almost every dictionary of chữ Nôm sorts entries using the *quốc ngữ* reading of each character. Where there are collisions or in appendices using stroke-based schemes, secondary orderings are based on total strokes or Kang Xi radical-stroke. In these stroke based orderings, *cá-nháy* are treated as additional strokes. *Nháy* is assigned 2 strokes and *cá* is assigned 3. This is illustrated in figure 6 below.

Names and Properties

As noted above, the properties of the *cá-nháy* are similar to the IDEOGRAPHIC TONE MARKs, so, ideally they would be placed within the same CJK Symbols and Punctuation block. Unfortunately, there is no longer space there, so an alternative would be to place them in the newer Ideographic Smbols and Punctuation block (16FE0 - 16FFF). The most salient property of the *cá-nháy* is their role as combining marks. The proposed properties reflect this.

16FF0;VIETNAMESE ALTERNATE READING MARK CA;Mc;6;L;;;;;N;;;;;

16FF1;VIETNAMESE ALTERNATE READING MARK NHAY;Mc;6;L;;;;;N;;;;;

LineBreak.txt

16FF0;CM   # Mc   VIETNAMESE ALTERNATE READING MARK CA

16FF1;CM   # Mc   VIETNAMESE ALTERNATE READING MARK NHAY

Representative Glyphs

While there maybe be variations in specific texts, the glyphs used in the Nom Na Tong font are based on the most common shape and should be used in the code charts

16FF0   VIETNAMESE ALTERNATE READING MARK CA   个

16FF1   VIETNAMESE ALTERNATE READING MARK NHAY   〈

Bibliography

Đinh Phan Cư. *Làm Quen với Chữ Quốc Âm* [Become Familiar with the Vietnamese Script]. Hán Nôm Thái Hòa Collection. Hà Nội. 2016.

Hoàng Thị Ngọ. *Về hiện tượng "cá nháy" và Các "ký hiệu phụ" trong chữ Nôm* [On the *cá nháy* and the *diacritic markers* in Nom characters]. Paper given at The International Nôm Conference, November 12-13, 2004, The National Library of Vietnam, Hà Nội. 2004.

Nguyễn Quang H`ồng. *Xem lại vai trò của "cá" và "nháy" trông cấu tạo chữ Nôm.* [A review of the rôle of "cá" and "nháy" in the structure of Chữ Nôm]. In *40 năm đào tạo và Nghiên cứu Hán Nôm (1972-2012)*. Hà Nội. Nxb Đại Học Quốc Gia Hà Nội. 2013.

Nguyễn Quang Hông. *Tự Điển Chữ Nôm Dẫn Giải*. [Dictionary of Chữ Nôm with source citations]. Social Sciences Publishing House, Hà Nội, 2014

Nguyễn Tá Nhí. *Lôĩ đanh dấu cá trong chữ Nôm*, [The Use of the cá diacritic in the Nôm script], Tạp chí Hán Nôm (1987) 1:35-38

TCVN 5773-1993. *The Nôm 16-Bit Character Standard Code Set for Information Interchange*. Technical Committee for Information Technology of Vietnam. Hanoi 1993.

Vũ Văn Kính. *Bảng Tra Chữ Nôm Miền Nam*. Hội Non Ngữ Học. TP. H`ồChí Minh, 1994.

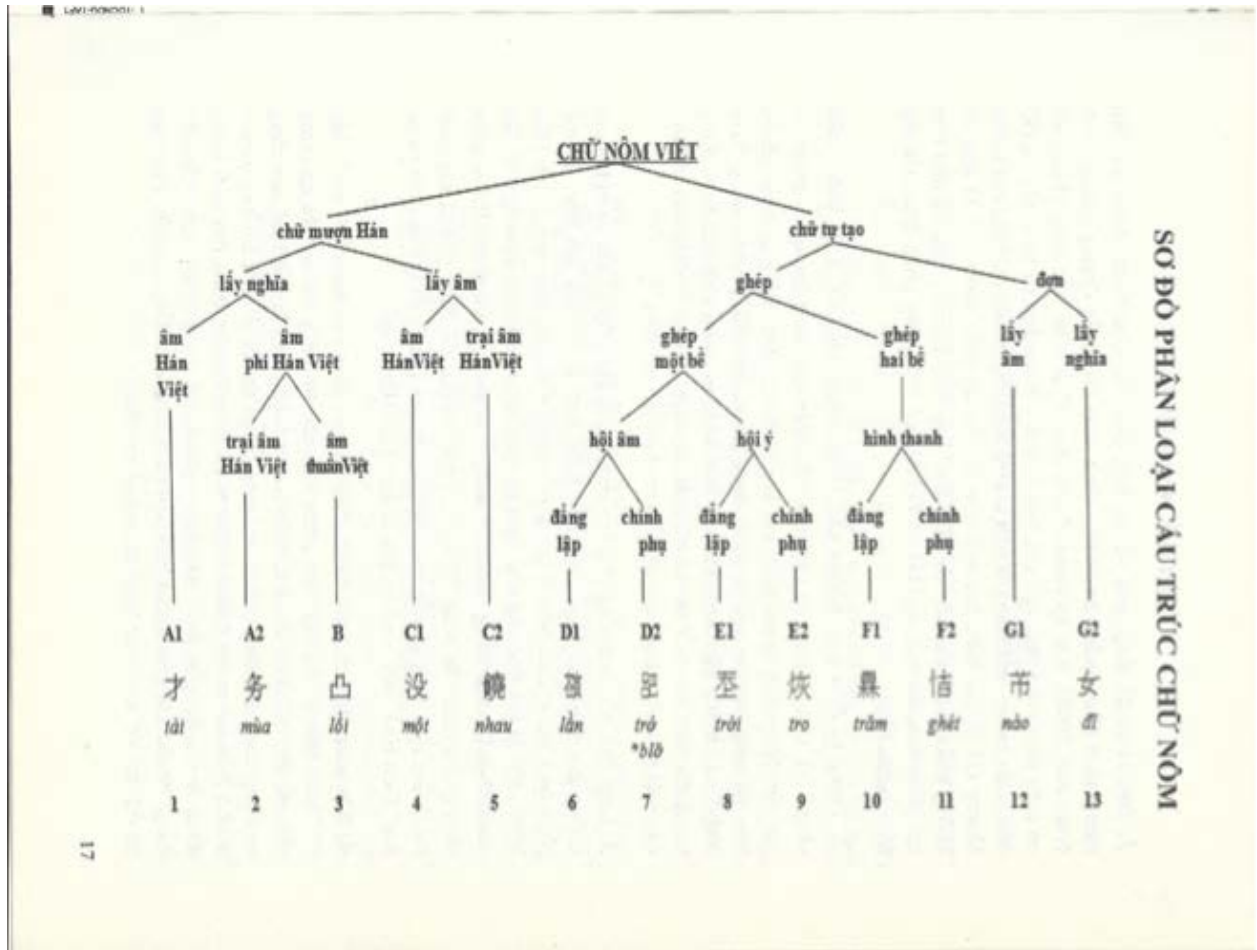**Figure 1 Structure of Nôm characters (H`ồng 2014, p 17)**

Figure 2. Examples of *cá* and *nháy* (Đinh 2014, p. 9)

Dấu Cá 个

Gọi là Dấu Cá 丫 vì nó chính là chữ Nho Cá 个. Dấu Cá 丫 được viết thêm vào ở bên phải chữ Nho.

Chữ Nôm	Nho	Ký hiệu	Đọc chệch
Cá 魛	Ngư 魚	个	
Mối 媒个	Môi 媒	个	Đọc chệch
Tiệc 席个	Tịch 席	个	Đọc chệch
Tuổi 歲个	Tuế 歲	个	Đọc chệch
Lỗi 磊个	Lỗi 磊	个	
Nợ 女个	Nữ 女	个	Đọc chệch
Tiếp 接个	Tiếp 接	个	
Tổn 損个	Tổn 損	个	

Dấu nháy ‹

Dấu Nháy 〈 cũng được viết thêm vào ở bên phải chữ Nho, thường là ký hiệu chỉnh âm.

Chữ Nôm	Nho	Dấu Nháy	Đọc chệch
Góa 寡‹	Qủa 寡	‹	Đọc chệch
Giành 爭‹	Tranh 爭	‹	Đọc chệch
Mày 眉‹	Mi 眉	‹	Đọc chệch
Mọc 木‹	Mộc 木	‹	Đọc chệch
Thà 時‹	Thì 時	‹	Đọc chệch

Có cả cách tạo chữ Nôm mới bằng cách thêm dấu nháy ‹ vào bên cạnh chữ Nôm.

Chữ Nôm	Nôm	Dấu nháy ‹	Đọc chệch
Lãn 嬾‹	Răn 嬾	‹	Đọc chệch
Mảnh 萌‹	Manh 萌	‹	Đọc chệch

Figure 3 TCVN 5773-1993 V+A171 *cá*

TCVN 5773-1993

Nôm Standard Code Set
Part 1

Table 1: COLLATING ORDER ACCORDING TO ISO/IEC 10646-1:1993

Code	Nôm	Vietnamese	Code	Nôm	Vietnamese	Code	Nôm	Vietnamese
V+A158	嗮	tấc	V+A174	㑔	luồn	V+A189	伽	nhờ
V+A159	朵	bủn	V+A175	帯	dưới	V+A18A	㓪	sệt
V+A15A	猫	chút	V+A176	梨	lè	V+A18B	恄	bẽ
V+A15B	宛	hoẻn	V+A177	齗	xuống	U+6091	怖	bỏ
V+A15C	伩	mọn	V+A178	㤱	ba	V+A18C	惷	cậy
V+A15D	乮	nhỏ	V+A179	㽬	dâng	V+A18D	㦢	cẩy
V+A15E	筊	xíu	V+A17A	蓮	trên	U+609C	惜	chường
V+A15F	尙	bé	V+A17B	杜	mất	V+A18E	惓	dãi
V+A160	買	mẩy	V+A17C	胡	quên	V+A18F	悢	dại
V+A161	贇	hèn	U+5FC6	忆	hắt	U+6087	悇	dơ
V+A162	舒	nghìn	V+A17D	仍	nải	U+608B	恰	lẫn
V+A163	兊	ngút	U+5FC7	协	sực	V+A190	仰	ngượng
V+A164	峀	núi	V+A17E	妠	nớ	U+60A7	悧	rượi
V+A165	嶀	đèo	V+A17F	仔	tệch	V+A191	忱	tấm
V+A166	嵿	vọi	U+5FCF	忓	thin	V+A192	悇	thoả
V+A167	㝮	mỏm	V+A180	忧	đăm	U+60CD	恾	căm
U+5D97	嶗	lao	V+A181	怺	thoải	V+A193	慌	hổ
V+A168	嫩	non	U+6011	怑	bản	U+60CF	惏	lầm
U+5E12	帒	đẩy	U+6032	怲	biếng	V+A194	悉	lòng
V+A169	崢	tranh	V+A182	体	bơn	V+A195	憐	lười
U+5E45	幅	bức	U+6026	怦	bừng	U+60C0	愉	luẩn
U+5E4D	帽	thao	U+601B	恛	dẫn	V+A196	恻	mừng
V+A16A	幱	mền	V+A183	怐	lo	U+60DD	憍	sượng
V+A16B	庉	đụn	U+6029	怩	náy	V+A197	悽	thiếp
V+A16C	庎	hè	V+A184	佣	rùng	V+A198	愉	tra
V+A16D	应	lụp	V+A185	㓸	thít	V+A199	愒	ủa
V+A16E	庴	chái	V+A186	惙	trót	U+60FC	愊	bẽn
V+A16F	弴	bắn	U+6044	恬	ghét	U+610A	愊	bực
V+A170	骒	khom	V+A187	㦐	ghiếc	V+A19A	愔	buồn
V+A171	个	cá	U+6045	恅	làu	V+A19B	惉	cáu
V+A172	舖	giuộc	V+A188	忛	lẫy	V+A19C	懂	chuộng
V+A173	吥	trụt	U+6051	恑	nguây	V+A19D	愂	dột

[12]

Figure 4. TCVN 5773-1993, V+A02F, *nháy*

Nôm Standard Code Set Part 1 — TCVN 5773-1993 — Table 1: COLLATING ORDER ACCORDING TO ISO/IEC 10646-1:1993

Code	Nôm	Vietnamese	Code	Nôm	Vietnamese	Code	Nôm	Vietnamese
V+A000	㥋	gộp	V+A013	儘	nghè	V+A030	齪	bấy
V+A001	仨	hai	V+A014	刨	vào	V+A031	吃	ốt
V+A002	莢	kép	V+A015	傪	tám	U+53ED	叭	bớt
U+4EC9	仉	kẻ	V+A016	汔	ngắt	V+A032	叮	léo
V+A003	企	trùm	V+A017	冻	mát	V+A033	㕧	nhấp
U+4F29	仗	nho	V+A018	潰	cóng	V+A034	代	dức
V+A004	伵	va	V+A019	刮	chểm	U+5403	吃	khật
V+A005	介	với	V+A01A	劇	phay	V+A035	吒	ngọt
U+4F4A	侅	bè	U+521F	刏	xẻo	V+A036	呃	rỉ
V+A006	伓	hùa	V+A01B	㖷	đẫn	U+5400	吀	xiên
U+4F32	伲	ni	V+A01C	剀	húi	V+A037	吶	ào
U+4F36	伶	rành	U+5256	剖	bõ	U+5431	哎	chê
U+4F35	伵	tớ	V+A01D	倒	xẻ	V+A038	吐	chửi
V+A007	偅	tốp	V+A01E	剧	hớt	U+5438	吸	cộp
U+4F85	伂	hời	V+A01F	㓹	khía	U+5432	呷	dặn
U+4FDA	俚	lái	V+A020	剌	mổ	U+5428	吨	dồn
V+A008	侱	vãi	V+A021	髙	cạo	V+A039	呻	diếu
V+A009	得	ngài	V+A022	削	gọt	V+A03A	呎	êu
V+A00A	侕	nhố	V+A023	劀	vạc	V+A03B	�	gáy
V+A00B	伩	nộm	V+A024	劤	rán	U+5474	呴	gâu
U+5045	偅	chống	V+A025	劲	gắng	U+542A	吡	hoé
V+A00C	媒	mụ	V+A026	努	khướt	V+A03C	吠	lóc
U+50A6	骨	cốt	V+A027	劫	rặn	V+A03D	吒	mau
V+A00D	納	nập	V+A028	劦	vâm	V+A03E	呀	ngát
U+5068	柴	thày	V+A029	舞	lử	U+5418	吤	ngỏ
V+A00E	絔	cấp	U+532C	匚	tủ	V+A03F	吭	ngoen
U+50E4	僤	dạn	V+A02A	逊	chục	U+5440	呀	nhá
V+A00F	僢	xúm	V+A02B	卧	bói	V+A040	吶	nhẵng
U+5100	儀	nghi	V+A02C	㕑	mái	V+A041	吐	nhằm
V+A010	偔	thợ	V+A02D	尬	chín	U+5450	呐	nói
V+A011	儡	trỗi	V+A02E	珍	dứa	V+A042	吨	uống
V+A012	傾	đòi	V+A02F	‹	nháy	U+5427	吧	và

[7]

Figure 5, Additional marks (Vũ, 1994. p 11)

Figure 6. Collation examples (Hồng, 2014, p. 2259)

2259

A. Administrative

1. **Title:** **Proposal to Encode Two Vietnamese Alternate Reading Marks**
2. Requester's name: *Lee Collins, Ngô Thanh Nhàn*
3. Requester type (Member body/Liaison/Individual contribution): *Liason*
4. Submission date: *4 October 2017*
5. Requester's reference (if applicable):
6. Choose one of the following:
    This is a complete proposal: *Yes*
    (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
    a. This proposal is for a new script (set of characters): *No*
       Proposed name of script:
    b. The proposal is for addition of character(s) to an existing block: *Yes*
       Name of the existing block: Ideographic Smbols and Punctuation block (16FE0 - 16FFF)

2. Number of characters in proposal: *2*

3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary		B.1-Specialized (small collection)	x B.2-Specialized (large collection)
C-Major extinct		D-Attested extinct	E-Minor extinct
F-Archaic Hieroglyphic or Ideographic			G-Obscure or questionable usage symbols

4. Is a repertoire including character names provided? *Ýe*
    a. If YES, are the names in accordance with the "character naming guidelines"
       in Annex L of P&P document? *Yes*
    b. Are the character shapes attached in a legible form suitable for review? *Yes*

5. Fonts related:
    a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
       Vietnamese Nôm Preservation Foundation
    b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
       Vietnamese Nôm Preservation Foundation, www.nomfoundation.org

6. References:
    a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
    b. Are published examples of use (such as samples from newspapers, magazines, or other sources)
       of proposed characters attached? *Yes*

7. Special encoding issues:
    Does the proposal address other aspects of character data processing (if applicable) such as input,
    presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *Yes*

8. Additional Information:
Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (

---

[9] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	*No*
If YES explain	
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	*Yes*
If YES, with whom?	*Scholars at Han-Nom institute in Hanoi, users active in Facebook group* Ủy Ban Phục Sinh Hán Nôm Việt Nam, etc.
If YES, available relevant documents:	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	*Yes*
Reference:	*Both scholarly and popular*
4. The context of use for the proposed characters (type of use; common or rare)	*Rare*
Reference:	
5. Are the proposed characters in current use by the user community?	Yes
If YES, where? Reference:	*See proposal*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	*No*
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	*Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	*Yes*
If YES, is a rationale for its inclusion provided?	
If YES, reference:	*Yes, but the proposed characters are combining marks, not stand alone characters*
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	*No*
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?	*Yes*
If YES, is a rationale for its inclusion provided?	*Yes*
If YES, reference:	*See proposal*
11. Does the proposal include use of combining characters and/or use of composite sequences?	*Yes*
If YES, is a rationale for such use provided?	
If YES, reference:	*Yes, see proposal*
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	*Yes*
If YES, reference:	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	*No*
If YES, describe in detail (include attachment if necessary)	
13. Does the proposal contain any Ideographic compatibility characters?	*No*
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	