A REVIEW OF DICTIONARY INDEXING AND LOOKUP METHODS FOR FOR IDEOGRAPHIC SCRIPTS IN COMPUTER

by
Ngô Thanh Nhàn
New York University
Courant Institute of Mathematical Sciences

Paper presented at
The First International Conference on Vietnamese Studies
July 14 - 17, 1998
Hanoi, Vietnam

ABSTRACT

The paper proposes a linguistic approach to indexing written ideographic syllables using internal regularities of their graphic representation. This approach allows one to break down syllables (written in squares) into orthographic units and their associated graphic operators. Stringing out these orthographic units and operators, the approach allows one to regenerate the known repertoire of ideographic written syllables, and possible but non-existing ones in the language. This method yields a more flexible indexing and dictionary lookup, and it also arrives at a complete and simple representation closest to the characteristics of ideographic syllables known and learned by native speakers.

INTRODUCTION

Computer character encoding for ideographic script is a relatively new international effort to include more than a third of human kind in the information technology arena for the next century. The East Asian countries and regions including China, Taiwan, Hong Kong, Korea, Japan, Vietnam, and Singapore... are getting together under the Ideographic Raporteur Group (IRG) chartered by the International Organization for Standardization, code name ISO/IEC 10646 – ISO/IEC JTC1/ SC2/ WG2/IRG. Working side by side with the IRG is the Unicode Consortium.

Vietnam joined this international effort to revive the interests in 喃 Nôm, an ideographic script that was used from the 10th Century until the beginning of the 1920's. The demise of Nôm was political. There is evidence that the French colonial regime had a plan to eradicate the Nôm script. Thus, one of the biggest losses in Vietnam during this century is Nôm. Within less than 50 years, the

number of people who could read and write in Nôm was reduced to less than 10. In the meantime, voluminous historical and literary documents continue to rot as the legacy of war continues to exact its toll. The revival of Nôm is not just a scholarly labor of love, it is a matter of protection of the heritage of a people.

The causes of lookup problems in ideographic dictionaries (we can consider this type of dictionaries as ordered lists of all known graphic symbols in these scripts) center around three contradicting concepts: the character, the syllable, and the ideograph. The term *ideographic character* is self-contradicting. This paper begins to address these problems on a theoretical level and proposes a solution based on the internal graphic consistency of ideographic repertoire.

What is called a character in ideographic scripts is actually a syllable. The term computer character encoding assumes that the basic element of encoding is a character ("a unit of information used for the organization, control, or representation of textual data" [1]). An ideographic syllable is currently incorrectly assumed to be a character, which puts it on an equal footing with a latin letter of the alphabet – reminding us of a *phoneme.* Stating that each syllable contains more than one character (unit of information) is an important first step.

The term *ideograph* – representing ideas (or meaning) with graphic symbols – exposes the internal arbitrariness in the relationship between meanings and graphic symbols. This assumption leads obviously to inconsistent and inadequate representation of ideographs. Thus, one of the chief methods of compiling syllable dictionaries (*tự điển*) of ideographic scripts is the KangXi Dictionary [2] (*Khang Hi tự điển*) and the like (such as the famous *Thuyết văn giải tự* "Deconstruction of Syllables", *Hán ngữ đại tự điển* "The Grand Chinese Dictionary", *Tân Hoa tự điển* "The New Chinese Dictionary", *Trung Hoa tân tự điển* "The Trung Hoa New Dictionary", etc.) in which each written syllable, 𡨸 *chữ* or 字 *tự,* is ordered (indexed) according to its 部 *bộ* or *bộ đầu* 部首 ("radical") and the number of strokes of its remaining part. *Bộ* ("radical") is a graphic segment that carries the "meaning" of the syllable. Each stroke is a continuous convenient movement of the brush tip in hand writing. We call this method collectively, the KangXi method.

The KangXi method rarely leads a lookup directly to the target syllable. It leads instead to a list of syllables that have the same number of strokes. The users have to visually identify the target. The word *bộ* should be literally translated as "a class, sort, genus", not "meaning" or "radical". The word *bộ* reminds us of the *morpheme,* but only graphically, because *bộ* does not exist in spoken language. It is in fact the major index of the written syllable. Complaints about this method are voluminous, such as inconsistencies in the number of radicals in each dictionary (214 in KangXi Dictionary, 540 in *Thuyết văn giải tự* "Deconstruction of Syllables", 200 in *Hán ngữ đại tự điển* "The Grand Chinese Dictionary", 189 in *Tân Hoa tự điển* "The New Chinese Dictionary", etc.), the choice of which radical a syllable belongs to, or how

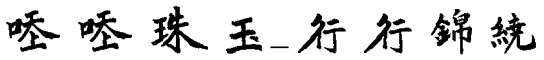
many strokes each syllable has, etc. I will not repeat these here. This method causes problem in automatic dictionary lookup and retrieval of syllables in an ideographic database, etc.

The KangXi method leads dictionary compilers to index new syllabic symbols with new radicals into the existing 214 *bộ*. It further leads to an untenable assumption that each written syllable must either be one of the fixed number of radicals, or containing one of those radicals. This obviously increases the arbitrariness of the dictionary compiling and retrieval (lookup) processes.

The mixed KangXi and latin-transcription method being used today does not address the inherent problems and inconsistencies of the KangXi method, rather, it increases the complexity of the lookup procedures. The users (or retrieval procedures) have to know the version of the latin-transcription of the standard dialect well to retrieve, not the target syllable, but a set of syllables having the same transcription – hoping that the intersection of the two non-unique solutions leads to a unique one. Of course, from experience, this is rarely true.

SOME THEORETICAL BACKGROUND

The *Tale of Kiều* (*Truyện Kiều*) has this verse written in Nôm:



transcribed into quốc ngữ, the current national latin-based standard script, as

> *Lời lời châu ngọc – hàng hàng gấm thêu*

and translated into English as

> each and every utterance is a pearl
> each and every phrase is a brocade.

Vietnamese is traditionally described such that when we speak, we utter one syllable at a time. Syllables string together into a phrase, and phrases string together into speech, like a string of pearls. A well-spoken lecture is compared to a brocade of woven strings of pearls. This observation is not far from today's basic theory of linguistics: the syllabic phonology and the property of *linearity* of language. Likewise, when we write, we can only write one stroke at a time. These strokes form phrases and texts, preserving the linearity of speech.

Language also has properties of *systematicity* and *universality*. We can think of these as properties of language which allow just any person to learn rather effortlessly.

We can think of a written language as a system of symbols representing human utterances. We note in passing that no system of writing (or phonetic transcription) can fully represent speech.

Vietnamese is a Mon-Khmer language of the Austro-Asiatic group. In Vietnamese, each spoken syllable, 喈 *tiếng*, is transcribed into one or more written syllables, or 𡨸 *chữ* (or 字 *tự*). This concept of *chữ* is preserved in both quốc ngữ and 喃 Nôm. Each written syllable, *chữ* – representing one or more *tiếng*, spoken syllables – is then written separated from each other. In quốc ngữ, as well as in Nôm, each *chữ* is bounded by delimiters, such as blanks, commas, periods, hyphens, question marks, quotation marks, exclamation marks, etc. A Vietnamese word consists of a positive integer number of syllables (*tiếng* or *chữ*), for example, *bút* ("pen"), *đồng hồ* ("watch, clock"), *nhà cửa* ("house"), *xe hơi* ("car"), *ô-tô* ("automobile, car"), ...

Some properties of ***quốc ngữ***

The quốc ngữ alphabet can be said to have two subsets: one is called orthographic (base character), and one tonal. The orthographic alphabet has 17 consonants and 12 vowels. The tonal alphabet has 6 tone marks. For example, the Vietnamese syllable *thấm* is spelled, out loud, as

â m[ờ] âm – th[ờ] âm thâm sắc thấm

We note that the two consonants *t* and *h* form one unit – the initial consonant cluster, *th[ờ]*. The rhyme *âm* is spelled first, then the segmental syllable *thâm*, then the tone *sắc*. We say that *â*, tone *sắc*, *t*, *h*, and *m* are *orthographic units*, and *ấ*, *t*, *h*, and *m* are *orthographic elements* [3]. An orthographic unit corresponds to one letter of the alphabet in quốc ngữ and one standard computer code point [4]. The tone *sắc* (as well as other tones in Vietnamese) is considered a *combining character* – a character that graphically combines with the preceding base character. [1]

In quốc ngữ, or any other latin-based scripts, there are two inherent assumptions in its display:

- each character is "housed" in an imaginary rectangular box (called a *cursor*);
- after a cursor is filled, the cursor moves right and waits for the next entry. The writing convention *left-to-right* has become the "norm" for all language scripts. This directional assumption creates a seemingly unnatural situation for quốc ngữ: when a tone is keyed in last in the syllable, like *sắc* after *thâm*, the tone mark "travels" backward, bypasses *m* and lands on *â*, to create *thấm*. [5]

We note that users of a latin-based script are allowed to combine letters of the alphabet at will without having to be concerned about whether the outcome is meaningful in the language or not. For example, we can create a string such as "*ccddeeff*" without editing intervention from the system. This means that defining letters of the alphabet (orthographic units) allows users to produce more than just existing words or syllables in a language.

Classifiers, radicals and morphemes

We say the nouns *con, tờ, hòn, nước, sợi, cái, cây, cuốn*, etc. are classifiers in the Vietnamese noun phrases such as *con dao* ("a knife"), *tờ giấy* ("a piece of paper"), *hòn đá* ("a rock"), *nước đá* ("ice"), *cái đá* ("a kick"), *sợi chỉ* ("a thread"), *cây kim* ("a needle"), *cuốn sách* ("a book"), respectively. These classifiers – for examples, *con* (classifier of animate objects), *tờ* (classifier of sheet-like objects), *hòn* (classifier of lump objects), *nước* (classifier of liquid objects), *sợi* (classifier of string-like objects), *cái* (classifier of inanimate objects), *cây* (classifier of stick-like objects), *cuốn* (classifier of scroll-like objects),… – follow some arbitrary convention in Vietnamese with respect to nouns. In Nôm, many syllables are believed to carry internal morphological classifiers, others do not. An internal morphological classifier in a Nôm syllable is not a morpheme (because it does not exist phonetically). It is graphically similar to, but not the same as, a combining mark, and is traditionally called *bộ* ("radical"). It is widely understood as the "meaning" part of the syllable. One study of how Nôm syllables use *bộ* can be found in Lê Văn Quán [6].

In *cho hắn một ... đá*,

- if classifier *cái* replaces "…", the phrase *cho hắn một cái đá* means "give him a kick";
- if classifier *hòn* replaces "…", the phrase *cho hắn một hòn đá* means "give him a rock."

In Nôm, however, it is clear whether *đá* is  a rock, with *bộ thạch* (classifier of rock), or *đá* is  a kick with *bộ túc* (classifier of foot or foot action). We say that in Nôm, classifiers usually agree (redundantly) in syntax and morphology. That is, *cái* appears to agree with *đá* with *bộ túc* , *hòn* appears to agree with *đá* with *bộ thạch* .

The syllable *đá* in the phrase, *đá lông nheo* ("beat the eyelashes"), can be conventionally classified either  with *bộ mao* (classifier of hair, feather, fur), or  with *bộ mục* (classifier of eye), or  with *bộ tiêu* (classifier of long human hair), etc. However,  with *bộ túc* is preferred, although *bộ túc* (classifier of foot or foot action) does not agree with *lông*  with *bộ mao* (classifier of hair) in  *lông nheo*. We say that, morphological classifiers, or *bộ*,

do not consistently agree with syntactic classifiers. Their correspondence is arbitrary. Furthermore, since *bộ* is a classifier, the system of classifiers of Han script and that of Nôm for Vietnamese do not correspond to each other either. As a result, indexing based on any of the Han systems of *bộ* (at least, according to the KangXi method) is neither precise, nor complete. The study of Nôm proper [7] and [8] shows that there are many *bộ* not found in the 214 *bộ* of the KangXi Dictionary. The same results were found in Korean, Japanese and even in Chinese proper.

Thus, using classifiers as an independent indexing system will not be consistent with respect to the graphic description (hence, indexing, retrieval, lookup,…) of ideographic syllables.

Each syllable in an ideographic script is written in an imaginary square. Like quốc ngữ, each *chữ* in Nôm is written between delimiters. And like quốc ngữ, each *chữ* has regularly recognizable pieces. In quốc ngữ, as shown above, such regularly recognizable pieces are called letters of the alphabet, or orthographic units. Thus, our problem in Nôm (as well as other ideographic scripts) is to identify these orthographic units. In quốc ngữ, letters of the alphabet are strung out from left to right (and stacking for tones) with a clear "syntax" of *chữ* uniquely identifiable [5]. Our problem in Nôm, likewise, is to identify the "syntax" of *chữ* that is uniquely defined or retrieved as intended.

Internal regularity of ideographic scripts

In Nôm, the syllables 肸 *mập,* 腧 *ỏng,* 肹 *phị,* 肭 *nục,* 腠 *béo,* 膿 *bọng,* 朦 *mẩy*,…(various degrees of obesity) have one element in common, 月 (called *nhục* – classifier of flesh or meat), see TCVN 5773:1993. "Spelling" in Nôm gives us a clue to identifying graphic units and their order in the mind of native speakers. Thus, when we "spell" 跺 *đá*, we say

write 足 *túc* on the left, 多 *đa* on the right (of the syllable square).

When we "spell" 劫 *kép* "compound", we say

write 二 *nhị* above, and 刼 *kiếp* below,

and when we "spell" 刼 *kiếp* "a life span", we say

write 去 *khứ* on the left, and 刃 *nhận* on the right.

So that, we can also "spell" 劫 *kép* "compound" as

write 二 *nhị* on top, 去 *khứ* below left, and 刃 *nhận* below right.

The following rules are taught when one learns the proper way to write ideographic syllables:

(1) first top, then bottom;
first left, then right; and
first outside, then inside.

The choice of which rule applies when, and their orders of application, seem to depend on the identification of regular "elements" of the syllable in the square. In fact, native speakers have no problem in identifying units in a syllable and how they are ordered in a square. The fact that each of the above rules is dualistic in nature allows us to algorithmically break the syllable into a string of orthographic units, strung out linearly from left to right. Word play (*chơi chữ*) shows evidence of this observation. For example, in a story I learned, a woman, named Phấn, introduced herself to her future husband, by spelling her name in a 5 syllable verse:

八刀 分 米 粉
bát đao phân mễ phấn
(literally, "eight broad knives divide the grain(s) into powder")

and he, named Chung, replied in perfect poetic counterpoint:

千 里重 金 鍾
thiên lý trọng kim chung
("[from] thousands of miles [away] [I come because] I value family life")

Her "spelling" is "八 over 刀 makes 分, add 米 [in front of 分] you get 粉".
His "spelling" is "千 over 里 makes 重, add 金 [in front of 重] you get 鍾."

We say that the system of writing in Nôm (or any other ideographic script) has regularly recognizable units known to (learned by) native speakers, and these units are ordered linearly and uniquely in a square.

ORTHOGRAPHIC UNIT AND STRING

From observation of handwriting behavior, we note that each ideographic syllable can be uniquely represented linearly by a number of orthographic units and their relative ordering operators (i.e. according to (1), top to bottom, left to right, and outside to inside).

The Chinese Delegation to the tenth IRG Meeting in Ho Chi Minh City in December 1997, proposed a set of BNF sequencing rules (named *ideographic structure sequence*) of idiographic units and their position operators (named *ideographic structure characters* – cf. Document No. ISO/IEC JTC1/SC2/WG2/IRG N518, N523 and N524). The proposal has 12 operators, visualized by slotted squares (⿴, ⿲, ⿳,

⿱, ⿴, ⿵, ⿶, ⿷, ⿸, ⿹, ⿺, ⿻), and an unknown number of orthographic units (radicals, ideograph components, and coded ideographs).

Mr Zhang Zhoucai (China) and Dr Lu Chin (HongKong) reported at the sixth IRG Meeting in Cupertino (USA), in February, 1996 (Document No. IRG N223), that the syllables which are composed of two elements occupy 96% of the syllables in KangXi Dictionary: 24% (around 12,000 syllables) ordered ⿱ *above-to-below*, 65% (around 32,000 syllables) ordered ⿰ *left-to-right*, 3.6% (around 1,800 syllables) ordered ⿸ *right-down encompass*, and 3.4% (around 1,700 syllables) ordered ⿺ *right-up encompass*.

Ngô Thế Lân of the Hán Nôm Institute, in an unpublished report (1996), found almost the same results while deconstructing the 1,770 Nôm proper syllables in TCVN 5773:1993. 9 of the 12 operators were found, among which operators No. 1 ⿰ (81%), No. 2 ⿱ (11%), No. 9 ⿸ (3%) and No. B ⿺ (4%) occupy an absolute 99.34%. The breakdowns are as follows:

No.	Op. No.	Operator	Description	Frequency
1.	0		* Losing one stroke	4
2.	1	⿰	Left-To-Right	1,434
3.	2	⿱	Above-To-Below	199
4.	3	⿲	Left-Middle-Right	** 0
5.	4	⿳	Above-Middle-Below	** 0
6.	5	⿴	Overall Around	2
7.	6	⿵	Down-To Encompass	2
8.	7	⿶	Up-To Encompass	** 0
9.	8	⿷	Right-To Encompass	1
10.		⿸	* Left-To Encompass	
11.	9	⿸	Right-Down Encompass	56
12.	A	⿹	Left-Down Encompass	2
13.		⿺	* Left-Up Encompass	
14.	B	⿺	Right-Up Encompass	70
15.	C	⿻	Embedded	** 0
		Total		1,770

* non-existing pattern in IRG N523.
** non-existing patterns in TCVN 5773:1993.

From a quick study of 2,357 syllables in the TCVN 5773:1993 database using two combining orthographic units to form a syllable (*chữ*), we found 287 first argument and 1,155 second argument graphemes, or a total of 1,321 units (121 appear in both

positions). 2 items, 个 *cá* and 𠂉 *nháy,* act like combining marks. 4 syllables – 共 *khệnh,* 共 *khạng,* 其 *khề,* 其 *khà* – belonging to pattern No. 0 – are not productive.

C C Hsieh, et al. [9] reported that

> "Components form a finite set. It is a closed set of approximately 1,200 finite elements. The rules of composition of a character [*read*, syllable] is very complicated from a traditional linguistical viewpoint,..."

These initial findings show that by defining a *hypothetical* set of 15 operators, shown above, and by using these operators to cut existing syllables into recognizable pieces (in Nôm and KangXi), we can verify the existence and productivity of each operator by its frequency of application. In the above exercise, 4 operators, No. 1 ⿲, No. 2 ⿳, No. 9 ⿵ and No. B ⿴ are productive both in KangXi and in Nôm. Furthermore, these operators yield a set of halves that are regular throughout the ideographic repertoire. Note that two operators are non-applicable, while operators No. 0 and No. C ⧉ are not functionally definable cuts.

Thus, it is desirable to require that binary operators must be recoverable. This means that if cuts performed on an ideographic syllable repertoire yield a certain number of halves, these halves must be recombinable by the same operators to yield the original syllables.

Logically, there are only 3 basic operators acting as binary cuts:

1. No. 1 ⿲ *left-right* (of which No. 3 ⿳ is a special case),
2. No. 2 ⿳ *top-bottom* (of which No. 4 ⿱ is a special case) and
3. No. 5 ⿴ *outside-inside* (of which Nos. 6 ⿵, 7 ⿶, 8 ⿷, 9 ⿵, A ⿸ and B ⿴ are special cases).

This conclusion is confirmed by (1).

The fact that syllables are being used graphically as parts of other syllables indicates that there are internal combinations of orthographic units. We find evidence that such internal combinations are binary, along the 3 basic operators cited above. For examples, 山 *sơn* ("mountain") is an independent syllable, and is also a part of the syllable 仙 *tiên* ("fairy"); 鳥 *điểu* ("bird") is an independent syllable, and 鳥 *điểu* is also a part of the syllable 鶏 *gà* ("chicken"). In TCVN 5773:1993 and TCVN 6056:1995 [10], we find (using prefix or Polish notation for operators:

- for 喱 *láy* ("reduplicate"), 俚 *lái* ("middleman"), 口 *khẩu* ("mouth"), 人 *nhân* ("person"), and 里 *lý* ("mile – land and nautical"), we can say they are formed by:

1 operator ⿴, and 3 orthographic units 口 *khẩu*, 亻 or 人 *nhân*, 里 *lý*, where

(a) the string "⿴ 亻 *nhân* 里 *lý*" → 俚 *lái*,

(b) the string "⿴ 口 *khẩu* ⿴ 亻 *nhân* 里 *lý*" → 喱 *láy*.

- for 糈 *bún* ("rice noodle"), 罘 *bốn* ("four"), 米 *mễ* ("rice"), 四 *tứ* ("four"), and 本 *bản* ("root"), we can say they are formed by

  2 operators ⿴ and ⿱, and 3 orthographic units 米 *mễ*, 四 *tứ*, 本 *bản*, where

  (c) the string "⿱ 四 *tứ* 本 *bản*" → 罘 *bốn*,

  (d) the string "⿴ 米 *mễ* ⿱ 四 *tứ* 本 *bản*" → 糈 *bún*.

- for 霓 *ngút* ("black, as in cloud"), 焆 *ngút* ("smoky"), 宊 *ngút* ("lofty, as in peak"), 雨 *vũ* ("rain"), 火 *hoả* ("fire"), 山 *sơn* ("mountain"), and 兀 *ngột* ("high"), we can say they are formed by

  2 operators ⿴ and ⿱, and

  4 orthographic units 雨 *vũ*, 火 *hoả*, 山 *sơn*, 兀 *ngột*,

  where

  (e) the string "⿱ 雨 *vũ* 兀 *ngột*" → 霓 *ngút*,

  (f) the string "⿱ 山 *sơn* 兀 *ngột*" → 宊 *ngút*,

  (g) the string "⿴ 火 *hoả* ⿱ 山 *sơn* 兀 *ngột*" → 焆 *ngút*.

- for 蹺 *nhào* ("tumble down"), 鼜 *nhiều* ("plenty"), 足 *túc* ("foot"), 多 *đa* ("many, much"), and 堯 *nhiêu*, *nghiêu* ("high, King Nghiêu"), we can say they are formed by

  2 operators ⿴ and ⿻, and

  3 orthographic units 足 or 𧾷 *túc*, 多 *đa*, 堯 *nhiêu;*

  where

  (h) the string "⿻ 堯 *nhiêu* 多 *đa*" → 鼜 *nhiều*,

  (i) the string "⿴ 𧾷 *túc* ⿻ 堯 *nhiêu* 多 *đa*" → 蹺 *nhào*.

We say that each of the strings (a) to (i) is the internal representation of the syllable after the "→" sign. An internal representation of a syllable, for our purpose here, is what is used for information interchange, storage, etc. in the computer. It is also the "spelling" of the syllable as we know it. For example, in (i), we say like Miss Phấn, "put 堯 *nhiêu* outside of 多 *đa* [operator No. 5 ⿻], you get 鼜 *nhiều*, and put 𧾷 *túc* in front of 鼜 *nhiều*[operator No. 1 ⿴], you get 蹺 *nhào*."

If we *successively* analyze (cut) the existing ideographic syllables into two parts, we will be able to arrive at the smallest graphic units – or orthographic units –

associated with their appropriate operators. The following BNF rule linearizes the observations (a)-(i) above and intuitive instruction in (1):

(2) <syllable> ::= <orthographic unit> [<syllable> | <orthographic unit>] .

where: each <syllable> is represented in an imaginary square shape.

We call each <syllable> an *orthographic element*. The linear concatenation (i.e. operator) of orthographic units to form an orthographic element <syllable> can be thought of as a predefined cursor frame which tells how two orthographic units are ordered and combine.

(3) There are 3 basic operators associated with the definition of each <syllable> of (2): No. 1 ⿲, No. 2 ⿳, and No. 5 ⿴. They can be designed in any order – prefix, infix or suffix.

We say a graphic operator is a control character which takes two arguments and graphically concatenates them in some specified manner. In this case, (3) defines three graphic control characters for ideographic scripts.

In the traditional ideographic textual layout (directionality) – syllables are printed top-down, right-to-left – thus, after a syllable frame is completed (initiated by a delimiter), the cursor moves down, and each carriage return means the cursor moves left to the top of the page. Within each syllable cursor box, defined by the BNF definition (2), orthographic units are fitted into their own subsquare spaces by the order that they are keyed in.

SOME PRACTICAL CONSIDERATIONS

1.
We shall not discuss rendering issues concerning well-formed combinations of orthographic units into syllables to be visible on display devices (on screen, in print, etc.). This is an efficiency issue for font implementation within the framework of internal representation of ideographic syllables proposed here.

The issue of full form vs. radical form needs further clarification. They are complementarily distributed with respect to shape. The rendered shape of an orthographic unit depends on its position in the syllable. For example, *nhân* ("human") has the form 亻 if it is on the left of a combination (the second argument of the operator ⿲), like 仙 *tiên* ("fairy"), and has the form 人 elsewhere, like 仝 *trùm* ("boss"), 以 *dĩ* ("by means of"), 囚 *tù* ("imprisoned"), 仄 *trắc* ("oblique"), or 人 *nhân* … We can write general graphic selection rules as follows:

(4) i. <orthographic unit> → *full-form* / <delimiter>+__+<delimiter>
ii. <orthographic unit> → *r-form* / <operator>+[<orthographic unit>+]__ +...
where: "+" string concatenation;
"__" position of the <orthographic unit> in "+" context.

Read, an orthographic unit has a graphic *full-form* if it is bounded by delimiters. An orthographic unit has a special graphic form (*r-form*) if it is the first or second argument of an operator. Thus, the rendering form of an orthographic unit depends on its position with respect to two elements: delimiter and operator.

A little more complicated example is *đao* ("knife"). It has the form 刂 if it is on the right of a combination (the second argument of the operator ⿰), like 利 *lợi* ("gain, sharp"), and has the form 刀 elsewhere (for examples, 分 *phân* ("to divide"), 刀 *đao* ("a knife")). But there are exceptions for the above rule, such as 刃 *nhận* ("sharp"), 切 *thiết* ("cut"), and 初 *sơ* ("original, new, initial"). However, note that the context condition of (4).ii. can be extended to cover these cases.

2.
Backward compatibility is inherent in the linguistic formalism described above because each "radical" is an orthographic unit (but not vice versa). This fact allows us to index ideographic syllables in the orders prescribed by previous dictionaries (i.e. 214 major indices for KangXi, 540 in *Thuyết văn giải tự*, 200 in *Hán ngữ đại tự điển*, 189 in *Tân Hoa tự điển*, etc.), but also allows us to successfully find the target syllable, i.e. each of which is identified by a single string of operators and orthographic units – not by their major indices (*bộ*) and number of strokes. From this point of view, the new method allows us to look for all syllables that contain a certain feature, such as orthographic units that have common features... For example, we can generate a list of all syllables (in a certain text) that have the orthographic unit 刀 and 刂 *đao* ("knife"); or a list of all syllables that have orthographic unit 刂 *đao* as second arguments of operator No. 1 ⿰.

3.
The basic problem in designing keyboard for ideographic scripts is the strict limit in the number of key strokes. The most popular keyboard today is the QWERTY 101. With thousands of orthographic units defined in a standard code table, how can we retrieve them using the current keyboard ? Note that with different data entry technology, such as latin "graffiti alphabet" for handwriting pens, the discussion may take a different turn. However, the discussion on ideographic strokes, be they key strokes or hand strokes, remains relevant.

The same internal regularity analysis used in defining orthographic units can also be used to address this problem. Our goals (or evaluation criteria) appear to be:

(5) Key strokes

1. identify the *least* number of key strokes, preferably less than 49 (shift and unshift alphabet keys, a-z and A-Z, minus 3 control operator keys in (3)), that include *basic strokes* and their associated *keystroke operators*;
2. identify the *smallest* sequence of basic strokes and keystroke operators that uniquely make up an existing orthographic unit, preferably the same number of strokes in the current "standard" stroke count; and
3. the keyboard entry sequence which is *closest to handwriting habit*, as described in (1) above, as well as the most convenient keyboard layout.

The criterion (5).3. is most important, because it is consistent with what native speakers have learned. It helps with literacy programs as well as computer popularization programs in these countries.

C C Hsieh et al. [9] give an example of the basic stroke set, with a note that the highest number of strokes encountered using this basic stroke set is 40:

S01	S02	S03	S04	S05	S06	S07	S08
							
S09	S10	S11	S12	S13	S14	S15	S16
							
S17	S18	S19	S20	S21	S22	S23	S24
							
S25	S26	S27	S28	S29	S30	S31	S32
							
S21A		S30A					
							

One example is given there: the syllable 灣 *loan* ("bay, gulf") is first broken down (by operator No. 1, 囗) into 氵 *thuỷ* ("water"), and 彎 *loan* ("to bend, curve"). 氵 *thuỷ* is then broken down into two S07 and one S08 strokes. 彎 *loan* is broken down (by operator No. 2, 曰) into 絲 *loan*? (archaic) and 弓 *cung* ("a bow"). The result is 29 strokes: 1 S08, 8 S07's, 6 S11's, 1 S15, 3 S05's, 4 S02's, 2 S01's, 2 S22's, 1 S13 and 1 S24 – as compared to 25 strokes in current "standard" stroke count. There is no formalism for composition of these strokes into intended orthographic units. I shall leave the subject for further investigation: we have proven that it is possible to design a keyboard for ideographic scripts satisfying the requirements in (5), and that the hint of a formalism for keystroke operators is, exactly, (1).

CONCLUSION

We have described *chữ* (or *tự*) in ideographic scripts as syllables written in an imaginary square. Each square is thus not comparable to a character. Each square is an independent syllable, bounded by delimiters, and decomposable recursively into regular units called orthographic units associated with three basic operators, ⿲, ⿳ and ⿴, as described in (2) and (3). An orthographic unit is the smallest unit of script, and each receives one computer code point. A basic operator is a control character having two arguments – each is an orthographic unit or an orthographic element – and arranging them in its given argument space. The decomposition, thus, transforms a written syllable into a linear string of operators and orthographic units. The decomposition is recoverable – i.e. decomposed elements are uniquely recomposable into syllables – in the manner similar to handwriting of native speakers.

We also propose the same decomposition method to break down orthographic units into basic strokes associated with a set of keystroke operators so that (a) the sequence of strokes and keystroke operators matches the writing behavior of native speakers, cf. (1), and (b) the strokes fit the current European keyboard setup. We propose a set of evaluation criteria (5) to choose the simplest and closest to native speakers' handwriting behavior.

We call our approach the linguistic approach – identifying internal graphic regularities within the repertoire of ideographic syllables with an aim to approximate native speakers' language intuition and behavior. We call our approach the two-tier approach: one for the information interchange (storage,...), and one for the keyboard entry. Obviously, it is possible to decompose the repertoire of ideographic syllables into basic strokes in a one-tier approach, the problem of composition of basic strokes into meaningful linguistic units, be it a character or a syllable, will increase in complexity and costs for information representation and information interchange.

REFERENCES

[1] *The Unicode Standard*, Version 2.0. Addison Wesley Developers Press. 1996.

[2] *KangXi Dictionary*. Trung Tân Library. 1981. Taiwan.

[3] J Đô, N T Nhàn, N Hoàng. 1992. A proposal for standard Vietnamese character encodings in a unified text processing framework. *Computer Standards & Interfaces* 14:3-10.

[4] Tiêu chuẩn Việt Nam. 1993. TCVN 5712:1993: *Information Technology – Vietnamese 8-bit standard code character set for information interchange* (VSCII). Hanoi, Vietnam.

[5] N T Nhàn. 1994. Some issues in automatic spell checking of Vietnamese written syllables without an associated spelling dictionary. Paper presented at *Tuân lê Tin hoc 4 -- The Fourth Biennual Technical Conference and Exhibition* (IW'94), Hô Chí Minh City, Viêt Nam. August 2-6, 1994.

[6] Lê Văn Quán. 1981. *Nghiên cứu về chữ Nôm*. Nhà xuất bản Khoa học Xã hội. Hanoi.

[7] Tiêu chuẩn Việt Nam. 1993. TCVN 5773: *Information Technology – The Nôm 16-bit character standard code set for information interchange – Chữ Nôm Việt*. Hanoi. Vietnam. 59 pp.

[8] Nguyễn Quang Xỉ & Vũ Văn Kính. 1971. *Tự điển chữ Nôm*. Trung tâm Học liệu. Saigon.

[9] C C Hsieh, C T Chang & J K T Huang. 1990. *On the formalization of glyph in the Chinese language.* A contribution to the AFII Meeting in Tokyo. Document No. ISO/IEC JTC1/SC2/WG2/IRG N292 (February, 1996 in Cupertino).

[10] Tiêu chuẩn Việt Nam. 1995. TCVN 6056: *Information Technology – The Nôm 16-bit character standard code set for information interchange – Chữ Nôm Hán*. Hanoi. Vietnam. 62 pp.

Dr Ngô Thanh Nhàn is a computational linguist associated with the New York University *Courant Institute of Mathematical Sciences,* and the *Aurum Language Systems.* He specializes in syntax and semantics of the medical language processing and automatic encoding of clinical narratives for a variety of Indo European languages. He is an expert in computer character encoding and was associated with Vietnam's efforts to standardize quốc ngữ, ideographic-based Nôm script and indic-based Chàm script since 1992. He is a board member of the *Brecht Forum,* an educational non-profit organization of New York State. He is a director of the *Vietnamese Heritage Institute,* a non-profit organization of the State of California, currently working on research projects and policy recommendations on environmental protection, agriculture, economy, and information technology of Vietnam. He is a founding member of *Peeling the Banana,* a New York Asian American performing arts collective.