BẢN XIN GÓP Ý

ĐƠN VỊ CHÍNH TÁ VÀ CÁC ĐẶC ĐIỂM CỦA TIẾNG VIỆT: CHỮ QUỐC NGỮ, HỆ LATINH, CHỮ NÔM, HỆ BIỂU Ý, VÀ UNICODE/ISO IEC 10646 ♣

Ngô Thanh Nhàn
Ban Chuẩn bị Sử dụng Bộ Mã chữ Việt theo Unicode/ISO 10646
Ngày 1 tháng 7 năm 2001

Tóm lược
Bài này dùng các tiêu chí của Unicode/ISO IEC 10646, như phổ quát, hiệu quả, đồng bộ và minh bạch, làm cơ sở cho tập mã đa ngữ Việt Nam. Các điểm chung của chữ quốc ngữ, chữ Nôm, chữ Chàm, chữ Thái, và các thứ chữ khác rõ ràng là ***tiếng***. Bài này cho thấy từ một kho chữ viết (một hệ thống chữ viết), ta rút ra được các đơn vị có hình dáng và có nghĩa nhỏ nhất của hệ thống chữ viết ấy, sao cho hệ thống luật tái tạo kho chữ ấy *đúng nhất* (tái tạo theo đặc thù của kho chữ này, nghĩa là ra tiếng Việt), *đầy đủ nhất* (tái tạo lại kho chữ ban đầu), *thông suốt nhất* (không thay đổi khi chuyển đi lại nhiều lần), và *đơn giản nhất* (hệ thống các ký hiệu và hệ luật kết hợp dễ thực hiện nhất). Ta gọi ký hiệu nhỏ nhất này là một ***đơn vị chính tả***. Các kho chữ trong tiếng Việt là ***kho ký hiệu tiếng*** và hoạt động đặc biệt của chúng trong tiếng Việt. Ở đây ta chỉ nói về chữ Nôm và chữ quốc ngữ hiện có mặt trong Unicode/ISO IEC 10646.

A. GIỚI THIỆU CHUNG

1. Trong bài *Cuộc đời sâu kín của Unicode—Lén xem chỗ nhược dưới bụng Unicode*,[1] bà Suzanne Topping ghi lại một cách dễ hiểu những bình phẩm của những người tham gia xây dựng và sử dụng Unicode và những người chống nó, để đưa đến kết luận, dù Unicode không làm mất đi các vấn đề qua cách quốc tế hoá, nhưng không ai bằng nó, và nó làm cho các vấn đề đặt ra thêm thú vị. Đa số vấn đề nằm ở nhóm chữ biểu ý CJKV (Trung-Nhật-Triều-Việt) do sự hiểu lầm chữ (tự) lâu đời của Âu Mỹ, mâu thuẫn giữa giải pháp mới cần kỹ thuật mới chưa phổ biến trong khi phải bảo vệ cái cũ lỗi thời nhưng có nhiều người dùng, v.v.

Bà Topping cũng bàn về mâu thuẫn tự nội của giải pháp mà Unicode gặp phải gọi là lộn xộn chuỗi mã tương đương *equivalency confusion*—ví dụ, bộ chuẩn có 4 mã, **o**, **ô**, **ố**, **dấu mũ** và **dấu sắc**, một con chữ phức như **ố** có nhiều hơn một cách tạo nó: (a) dùng 1 mã **ố** dựng sẵn, (b) tạo **ố** bằng hai mã: **ô** và **dấu sắc**, hoặc (c) tạo **ố** dùng ba mã: **o**, **dấu mũ** và **dấu sắc**. Tuy Unicode

♣ Xin cám ơn góp ý của Trần Lưu Chương, Ngô Trung Việt, (James) Đỗ Bá Phước, Vũ Quang Việt, Hà Dương Tuấn, Hồ Văn Tiến, Lê Phạm Ngưng Hương.

[1] Suzanne Topping, The secret life of Unicode—A peek at Unicode's soft underbelly, *IBM Developer Works* (http://www-106.ibm.com/developerworks/unicode/library/u-secret.html), 5/2001.

không khuyến khích dùng (a), nhưng đây có lẽ là vấn đề thảo luận sôi nổi nhất của chữ quốc ngữ khi dùng Unicode.

Bài này nêu ra một số vấn đề của chữ quốc ngữ, chữ Nôm và các thứ chữ âm tiết khác như chữ Chàm và chữ Thái, đưa thêm một số tiêu chí để đánh giá chúng theo yêu cầu của từng thứ tiếng dân tộc trong nước Việt Nam. Luận điểm chính là Unicode (hay bất kỳ một bộ mã nào) chỉ có giá trị khi nó đáp ứng được yêu cầu thông tin đúng, đầy đủ, thông suốt và đơn giản của một thứ tiếng nói cụ thể.

2. Chuẩn mã hoá ký tự (*character encoding*) công nghệ thông tin lấy tính phổ quát (*universal*), hiệu quả (*efficient*), đồng bộ (*uniform*) và minh bạch (*unambiguous*) làm mục tiêu chính.[2] Chúng ta dùng chuẩn Unicode Consortium 16-bit hay ISO IEC 10646 32-bit (ta gọi tắt là *Unicode*) có nhiều lợi thế hơn các chuẩn ký tự 8-bit trước đây, cơ bản là:

- Unicode có đủ chỗ chứa chữ quốc ngữ, chữ Nôm, chữ Chàm, chữ Thái, và nhiều thứ chữ khác của Việt Nam (*tính phổ quát*) trong cùng một bảng. Hệ luận: những công cụ tìm kiếm, sửa đổi, không cần phải phân biệt khoá tìm có phải là quốc ngữ hay chữ Nôm hay chữ Chàm, hay bất cứ thứ chữ gì khác trên thế giới (*tính đồng bộ*).
- Cho phép có dấu rời như 5 dấu thanh trong chữ quốc ngữ, các dấu nguyên âm trong các ngôn ngữ dân tộc, các dấu nguyên âm trong chữ Chàm, chữ Thái, v.v. có thể dùng chung mà không lẫn lộn (*tính minh bạch*).
- Cho phép ta viết một nửa chữ Nôm, một nửa chữ quốc ngữ, một nửa chữ Chàm, chữ Thái, v.v. mà vẫn tìm ra (*tính minh bạch*).

Unicode không có trách nhiệm làm riêng cho một thứ chữ viết của quốc gia nào. Nghĩa là Unicode cho phép một văn bản chứa nhiều loại chữ viết (tính đa ngữ, *multi-lingual*). Trong định nghĩa ký tự (*character* hay *code value*) của Unicode, một con chữ cái latinh, một chữ biểu ý hay chữ gốc ấn đều mang một mã—trong khi một chữ biểu ý hay một chữ gốc ấn là một âm tiết, tương đương với một chuỗi mã latinh.[3]

Như thế, vì chữ Nôm, chữ Chàm, chữ Thái và chữ quốc ngữ đều hiện diện trong Unicode, tôi coi đó là lợi thế ta cần nghiên cứu.

- Tiếng Việt có tính đơn tiết (*monosyllabic*). Mỗi âm tiết (*syllable*) tiếng Việt nói rời nhau ra. Ví dụ, từ "Việt Nam" có hai tiếng. Cấu tạo hình vị (*morpheme*—đơn vị nghĩa nhỏ nhất có dạng âm thanh), từ (*word*—đơn vị tạo câu nhỏ nhất, gồm một hay nhiều hình vị, cộng nghĩa thành nghĩa của từ, ví dụ, *ab* "không" + *normal* "bình thường" → *abnormal* "không bình thường"), ngữ đoạn (*phrase*), câu (*sentence*), v.v. trong tiếng Việt gồm ***một***

---

2 *The Unicode Standard*, Version 2.0 (1996). Addison-Wesley Developers Press. Trang 1-2.
- ***Universal***. The repertoire must be large enough to encompass all characters that were likely to be used in general text interchange, including those in major international, national, and industry character sets.
- ***Efficient***. Plain text, composed of a sequence of fixed-width characters, provides an extremely useful model because it is simple to parse; software does not have to maintain state, look for special escape sequences, or search forward or backward through text to identify characters.
- ***Uniform***. A fixed character code allows efficient sorting, searching, display and editing of text.
- ***Unambiguous***. Any given 16-bit value always represents the same character.

3 *Sách đã dẫn*, trang G-2.

số nguyên các âm tiết.[4] Tiếng Anh có hình vị số nhiều *–s* nhỏ hơn một âm tiết. Một hình vị hay một từ tiếng Việt nhỏ nhất là một âm tiết.

- Mỗi chữ Nôm, chữ Thái hay chữ Chàm phát âm thành một âm tiết—tương đương với một chuỗi con chữ quốc ngữ. Định nghĩa cấu hình một âm tiết của chữ quốc ngữ cho phép ta chuyển tiếng–chữ Nôm/Chàm/Thái–chữ quốc ngữ thành một hệ một–đối–một biểu diễn đầy đủ tiếng Việt.
- Trong bài này, do đó, chúng ta chọn âm tiết làm đơn vị ngữ cảnh nhỏ nhất, làm điểm chung cho chữ quốc ngữ, chữ Nôm, chữ Chàm và chữ Thái, v.v.

Riêng việc làm chuẩn công nghệ thông tin trong tiếng Việt phải tuân theo những chuẩn đã có như:

- chuẩn chữ viết quốc ngữ của Viện Ngôn ngữ, các từ điển tiếng Việt hiện đại, các sách giáo khoa hiện đại, như chuẩn các con chữ cái (xem các từ điển), chuẩn bỏ dấu thanh lên một âm tiết, chuẩn chính tả (còn lơ mơ), và tiếng nói chuẩn (nhắc đến nhiều trong ngành giáo dục, ít được ai nhắc đến trong công nghệ thông tin).
- người hành nghề CNTT khó viết được hệ tìm kiếm, sắp thứ tự, in đẹp, trình bày đẹp, hoặc nhập sửa văn bản, nếu không có chuẩn bỏ dấu thanh, chuẩn chính tả, v.v. ***theo hệ âm tiết***.
- khi có nghi vấn, thì các tập tục xung quanh việc sử dụng chữ viết tiếng Việt trong các hoạt động văn hoá—xã hội thường nhật và cuối cùng là chuẩn tiếng nói. Ví dụ, khi nói "ba", người lục sổ tìm chữ "ba", "Ba", "BA" và số 3. Khi nói "rượu", người lục sổ tìm bằng mắt các chữ "rịu", "dịu", "riệu", "rượu", "diệu", v.v. và một số cách đánh sai. Đây là các yêu cầu mà chuẩn mã ký tự CNTT phải chú ý đến ngay từ đầu.

Xưa nay, chúng ta nói chuẩn chữ viết hàm ý chuẩn tiếng nói. Đáo cùng, chuẩn tiếng nói là cơ bản nhất. Trong bài này, ta giả định có tiếng nói chuẩn của Việt Nam. Ta giả định giữa tiếng nói chuẩn và chữ quốc ngữ có quan hệ một-đối-một (dù trong chữ quốc ngữ và chữ Nôm có một số nhược điểm).[5] Ta coi đơn vị mô tả trong tiếng Việt chuẩn là tiếng (một âm tiết) và các hoạt động của nó trong toàn bộ ngôn ngữ.

B. ĐƠN VỊ CHÍNH TẢ

Ta gọi một đơn vị chính tả (*orthographic unit*) [6] là một đơn vị nhỏ nhất có hình dáng và có nghĩa của một hệ thống chữ viết. Một đơn vị chính tả được biểu thị bằng một mã ký tự chuẩn.

Định nghĩa này cho phép một đơn vị chính tả có tính trừu tượng, nhưng luôn luôn có hình dáng, ví dụ, đơn vị chính tả "a A …" là con chữ cái "a" trừu tượng, mang nhiều hình dáng khác nhau. Ví dụ, chữ Nôm ***trời*** có hai đơn vị chính tả, ***thiên*** trên và ***thượng*** dưới. Định nghĩa này buộc chúng ta nhận dấu thanh (huyền, sắc, nặng, hỏi, ngã) là đơn vị chính tả (một mã ký tự riêng),

---

4 Ngô Thanh Nhàn, *The syllabeme and patterns of word formation in Vietnamese* [Tiếng và các mẫu cấu tạo từ trong tiếng Việt]. Luận án tiến sĩ, Đại học New York. 1984. Abstract, Trang 1-2.

5 *Sách đã dẫn*. Phụ Lục A, Orthographic-Phonological Conversion [Chuyển đổi chính tả–âm vị], trang A1-A6.

6 A proposal for standard Vietnamese character encodings in a unified text processing framework, James Đỗ Bá Phước, Ngô Thanh Nhàn và Nguyễn Hoàng. *Computer Standards & Interfaces* 14 (1/1992):3-10.

trong khi đó, các dấu nguyên âm (mũ *circumflex*, trăng *breve*, râu *horn*, v.v.) không phải là đơn vị chính tả trong chữ quốc ngữ.

Nhóm từ "của một hệ thống chữ viết" trong định nghĩa trên cùng nghĩa với từ kho (*repertoire*) trong Unicode. Nghĩa là các đơn vị chính tả của chữ Nôm là hình dáng phân tích có nghĩa nhỏ nhất trong kho chữ Nôm mà thôi. Đơn vị chính tả của chữ quốc ngữ là hình dáng có nghĩa nhỏ nhất trong kho chữ quốc ngữ. Nó có nghĩa trong phân tích nội tại của một kho chữ. Ví dụ, dấu mũ (*circumflex*), trăng (*breve*), râu (*horn*), v.v. không có nghĩa trong chữ quốc ngữ, nhưng chúng có nghĩa trong tiếng Pháp, tiếng Tây-ban-nha, ti ếng Bồ-đào-nha, chẳng hạn.

1. Phương pháp luận ở đây bắt đầu bằng một kho chữ. Ai cũng phải làm thế. Từ kho chữ ấy, ta rút ra những bộ phận giống nhau nhỏ nhất, cắt tuần tự theo nhiều phương pháp khác nhau và theo dõi toàn bộ các quy trình cắt ấy. Từ một kho chữ có chiều dài nhất định, chúng ta luôn luôn có nhiều quy trình cắt khác nhau thành những đơn vị khác nhau. Mỗi hệ thống cắt cho ta một hệ đơn vị chính tả. Đảo ngược một hệ thống cắt,[7] ta có một hệ thống kết hợp riêng cho hệ thống đơn vị chính tả liên hệ. Tất cả các hệ thống kết hợp và đơn vị chính tả của chúng đều sản sinh ra cùng một kết quả (kho ban đầu) như ý.

Ví dụ, phân tích kho chữ quốc ngữ, ta có thể có 3 giải pháp (nhớ lại những ngày đầu của chương trình chuẩn hoá):

- Giải pháp dựng sẵn (*precomposed*): 72 nguyên âm, 20 phụ âm, cách ghép chữ đơn giản nhất: con trỏ chạy từ trái sang phải. Nhắc lại giải pháp hết sức thông minh 2 bộ phông hoa và thường.
- Giải pháp từ điển Việt Nam (nửa kết hợp): 12 nguyên âm, 5 dấu thanh, 20 phụ âm, cách ghép chữ đơn giản nhất: con trỏ chạy từ trái sang phải, 5 dấu thanh "múa" trên âm tiết. Cách này phục tùng chuẩn chữ quốc ngữ. Nhắc lại giải pháp biểu hiện TrueType Font hết sức thông minh: dấu thanh có chiều rộng 0.
- Giải pháp kết hợp (*decomposed*): 6 nguyên âm, 8 dấu (3 dấu nguyên âm, 5 dấu thanh), 20 phụ âm, cách ghép chữ đơn giản nhất: con trỏ chạy từ trái sang phải, các dấu "múa" trên âm tiết. Nhắc lại giải pháp biểu hiện TrueType Font hết sức thông minh: dấu có chiều rộng 0. Đáng chú ý là hai giải pháp trên im lặng dùng giải pháp này để làm bàn phím.

Unicode gồm cả 3 giải pháp cùng lúc và buộc phải dùng bảng chuỗi mã tương đương.

Do đó, chúng ta cần có những tiêu chí độc lập với tất cả các quy trình để đánh giá chúng.

Trong một thứ tiếng, chúng ta không bao giờ tìm được kho đầy đủ (ví dụ, kho chữ Hán-Nôm có thể ngày càng nhiều, nhưng không bao giờ đủ, kho tiếng Việt không bao giờ đầy đủ). Hệ các đơn vị chính tả và hệ kết hợp giúp chúng ta tìm ra đặc thù của một thứ chữ viết cho một thứ tiếng… nhờ đó chúng ta tiệm cận được đúng và đầy đủ của một thứ chữ viết (mà không cần khởi đầu bằng một kho đầy đủ).

[7] Có hệ thống cắt không đảo ngược được. Do đó, chúng ta thường chọn các cách cắt là những chương trình xử lý chuỗi (*string functions*) đảo ngược được.

Nếu không có tiêu chí sắp thứ tự đúng chữ quốc ngữ (các âm tiết), v.v., thì phương pháp dựng sẵn (*precomposed*) bằng hoặc hơn hẳn phương pháp kết hợp (*combining*) trong bối cảnh kỹ thuật 8-bit những năm trước đây. Nhưng khi có thêm các tiêu chí về hoạt động khác của ngôn ngữ, như sắp thứ tự, tìm kiếm, bỏ dấu thanh đúng chỗ, v.v. phương pháp kết hợp bắt đầu cho thấy tác dụng của nó… tuy nó đòi hỏi phải có kỹ thuật mới (kỹ thuật kết hợp) cho trình bày và in ấn.

Phương pháp tổ hợp hoàn toàn (coi các dấu nguyên âm là các đơn vị chính tả) tuy không đúng chuẩn từ điển chữ quốc ngữ trong tiếng Việt nhưng lại có lợi trong một số thứ tiếng dân tộc. Kèm theo việc phân tích kho chữ không cần nhìn vào tiếng nói đưa ta đến lựa chọn này. Vì chưa có một giải pháp chữ quốc ngữ nào sử dụng phương pháp này (trừ bàn phím), tôi xin để dành phân tích phương pháp này cho tương lai, vì chúng ta không dễ dàng lơ đi.

2. Thêm nữa, ta có thể gọi một đơn vị chính tả của chữ quốc ngữ là ký hiệu biểu thị một đơn vị chính âm (âm vị, *phoneme*) theo từ điển chuẩn. Đơn vị chính âm là một đơn vị trong tâm thức của người bản xứ. Ví dụ, vần là một đơn vị trong tâm thức của người nói tiếng Việt (có thể nói vần là đơn vị âm thanh phổ quát—mọi ngôn ngữ đều có thi ca, dùng vần điệu trong thi ca). Trong bài này chúng ta bắt đầu dùng đơn vị chính tả gần với đơn vị chính âm để cho thấy sự cần thiết phải tiệm cận chính tả và tiếng nói, cho đơn vị chính tả cái ta gọi là chứng cớ thực tế sự hiện hữu của nó (*physical evidence*) trong ngôn ngữ.

Ta có thể nghe/thấy được các đơn vị chính tả qua cách đánh vần của một dân tộc. Cùng một âm tiết /xem/, cách đánh vần chữ quốc ngữ cho ta cấu tạo nội tại và các đơn vị chính tả trong chữ quốc ngữ—khác với cách đánh vần chữ Nôm. Đánh vần là chỉ cho người khác cách viết đúng như ý mình.

> Chữ quốc ngữ: xem—**e mờ em xờ em xem** [8]
> Chữ Nôm: xem—**mục** bên trái, **chiêm** bên phải

Cách đánh vần chữ quốc ngữ cho ta các đơn vị chính tả: **e** **m**ờ **x**ờ, và các đơn vị cao hơn, **e**, **em** và **xem**. Đơn vị **em**, ta gọi là vần của đơn vị **xem** ta gọi là tiếng. Đặc điểm của cách đánh vần này là ta không bắt đầu từ trái sang phải, mà bắt đầu từ nguyên âm trung tâm **e** (đã là một tiếng), xong thêm **m** để làm thành vần **em** trước, xong mới cộng thêm phụ âm **x**, xong mới thêm dấu thanh, để thành tiếng **xem**.

Ngược lại, cách đánh vần trong chữ Nôm cho ta hai đơn vị chính tả: **mục** và **chiêm**. Đơn vị chính tả "mục" cho ta vùng nghĩa của chữ **xem**. Đơn vị chính tả "chiêm" cho ta vùng âm thanh của chữ **xem**.[9] Chữ Nôm vì đã có hơn 10 thế kỷ, vùng âm thanh ghi lại những chặng biến đổi âm thanh trong lịch sử phát triển tiếng Việt. Vùng nghĩa cho ta biết loại từ (*classifier*, còn gọi là *bộ*) của chữ **xem**.

---

[8] Chữ quốc ngữ có những chỗ "hơi" bất ngờ (do lịch sử để lại) như các nhóm phụ âm cuối *–ch*, và *–nh* khi phát âm thành /k/ **cờ** và /ng/ **ngờ**:
Chữ quốc ngữ: (tấp) tểnh—**ê nhờ ênh tờ ênh tênh hỏi tểnh**
Chữ Nôm: (tấp) tểnh—**tâm** trái, **tỉnh** phải.

[9] Xem thêm, Ngô Thanh Nhàn, *A review of dictionary indexing and lookup methods for ideographic scripts*, trình bày tại *Hội nghị Việt học Lần thứ nhất*. Hà Nội (14-17.7.1998), cf. http://www.cs.nyu.edu/~nhan/vsic98.pdf.

Một tiếng nói có hai thứ chữ viết theo hai hệ thống khác nhau mang cho tiếng Việt nhiều lợi thế.

C. TIẾNG VÀ CÁC YÊU CẦU CHUẨN CNTT

Tiếng là đơn vị mà chuẩn các chữ viết (Nôm, quốc ngữ, Chàm. Thái) và chuẩn tiếng nói gặp nhau. Không phải vô tình mà hai bên một chữ quốc ngữ và một chữ Nôm (một chữ Chàm hay một chữ Thái) có các dấu cách. Tiếng là âm tiết. Chữ (hay tự) là ký hiệu (hình vẽ) của tiếng. Như vậy, về mặt chữ viết, chúng ta chọn chữ làm một đơn vị nghiên cứu để mô tả tiếng là một đơn vị âm thanh.

Ở đây, ta chọn yêu cầu "đúng", "đầy đủ", "thông suốt", và "đơn giản" làm thước đo các giải pháp chuẩn ký tự.

- Yêu cầu "đúng" đòi hỏi mô tả được cái đặc thù của một thứ tiếng (bỏ dấu thanh ở đâu trên một chữ tiếng Việt ta vẫn tìm ra). Yêu cầu này hàm ý làm mạnh hơn tính hiệu quả của Unicode. Ở đây, chúng ta chọn biểu diễn đúng các thứ tiếng Việt, và gần nhất với cách viết tay và cách đánh vần. Người sử dụng tự nhiên với máy tính là niềm vui của người làm CNTT thay vì nó có nghĩa là người làm CNTT "cực khổ" hơn.
- Yêu cầu "đầy đủ" không chỉ tái tạo toàn bộ kho chữ tiếng Việt hiện có, mà còn dành chỗ cho những khả năng phát huy trong tương lai (vì tiếng nói luôn thay đổi)—tiếng địa phương, âm nói được nhưng không có nghĩa (ví dụ, trong ngành trình diễn, như *Benny Hill*, có khi diễn viên nói một tràng tiếng Anh nhưng chỉ gồm những tiếng—*syllables*—thuần Anh ngữ nhưng nhập lại thành vô nghĩa).
- Yêu cầu "thông suốt"—hình dáng và chức năng không thay đổi—đòi hỏi dữ liệu được bảo vệ đúng trong mọi đường truyền. Yêu cầu này và yêu cầu đơn giản hàm ý làm giảm thiểu các chuỗi ký tự tương đương.
- Yêu cầu "đơn giản" cho phép ta chọn giải pháp xử lý nhanh nhất. Ở đây dĩ nhiên ta không cần nói ra yêu cầu này, nhưng ý chúng tôi là các giải pháp mạnh dạn đưa ra kỹ thuật mới để luôn luôn đơn giản hoá quy trình.

Do đó, bài này chúng ta nghiên cứu về các đơn vị chính tả trong tiếng Việt sao cho chúng phản ánh đúng các đơn vị âm thanh và các hoạt động của chúng trong tiếng Việt.

Chữ quốc ngữ tiếng Việt gồm có:

1. 29 chữ cái, theo các từ điển hiện đại,

   a ă â b c d đ e ê g h i k l m n o ô ơ p q r s t u ư v x y

   4 chữ cái để ghi tiếng các dân tộc khác: **f, j, w, z**.
   và 5 dấu thanh, Ò (*huyền*), Ỏ (*hỏi*), Õ (*ngã*), Ó (*sắc*), Ọ (*nặng*), viết trên nguyên âm.

2. 16 nguyên âm, viết thành 14 nhóm chữ cái,

   a, ă, â, e, ê, i/y, ia/iê/ya/yê, o, ô, ơ, u, ua/uô, ư, ưa/ươ

ngắn	dài
ă	a
â	ơ
(ach, anh)	e
(êch, ênh)	ê
i, y	ia, iê, ya, yê
	o
	ô
u	ua, uô
ư	ưa, ươ

Xem, *The Syllabeme*… sách đã dẫn.

3. 24 phụ âm đầu (một phụ âm đầu tắc hầu, *glottal stop*, không có con chữ cái), viết thành 23 nhóm chữ cái, và 4 chữ cái cho tiếng dân tộc khác (viết trong ngoặc đơn),

   b, c/k/q, ch, d, đ, (f), g/gh, gi, h, (j), kh, l, m, n, nh, ng/ngh, p, ph, r, s, t, th, tr, v, (w), x, (z)

4. 1 bán nguyên âm đầu (tròn môi, **o** hay **u**): Ví dụ, khoan, khuynh, noãn, công-poanh, nguyễn, v.v. Chứng cớ bán nguyên âm này là một phần âm sắc của phụ âm đầu là nói lái <u>hoà lan</u> thành <u>hoàn la</u> (âm tròn môi o đi theo h).

5. 2 bán nguyên âm cuối (i, y, o, u), 6 phụ âm cuối (p, t, c/ch, m, n, ng/nh)

   i/y, o/u, p, t, c/ch, m, n, ng/nh

6. 6 thanh, viết bằng 5 dấu. Thanh ngang không mang dấu.

   Sự phân biệt bằng trắc, cao thấp, giúp chúng ta tái tạo cách nói lái (*đấu tranh*, *đánh trâu*, *tránh đâu*, *trâu đánh*, *tranh đấu*,…), lập từ láy (*trăng trắng*, *nho nhỏ*, *mằn mặn*, *vò võ*, v.v.), ngữ đoạn, …

	bằng	trắc	
cao	ngang (*đoản bình*)	sắc (*thượng*)	hỏi (*hồi*)
thấp	huyền (*trường bình*)	nặng (*hạ*)	ngã (*khứ*)

7. Một tiếng trong tiếng Việt gồm có một phụ âm đầu, một bán nguyên âm, một nguyên âm chính, một phụ âm hay bán nguyên âm cuối và một thanh.

tiếng				
thanh	**phụ âm**		**vần**	
thanh	phụ âm đầu	bán nguyên âm	nguyên âm chính	phụ âm / bán nguyên âm cuối
t	**P**	**W**	**V**	**C**

Trong lịch sử, một tiếng gồm một phụ âm (phụ âm đầu + bán nguyên âm tròn môi), một vần (nguyên âm chính + phụ âm/bán nguyên âm cuối) và một thanh. Một tiếng phải có ít nhất một thanh và một nguyên âm chính, các thành phần khác của tiếng xuất hiện theo các mẫu dưới đây. Sự phân biệt phụ âm, vần và thanh mô tả tiếng nói đầy đủ nhất.

a) tV
b) tWV
c) tVC
d) tWVC
e) tPV
f) tPWV
g) tPVC
h) tPWVC

Nói như thế thì một thanh và một nguyên âm chính đã lập thành một tiếng. Cặp vần+thanh là một tiếng. Phụ âm đầu, bán nguyên âm, phụ âm/bán nguyên âm cuối đều là phụ gia. Một phụ âm đầu không làm thành một tiếng. Vần là một tiếng.

Có một số luật kết hợp chuẩn giữa các đơn vị tiếng (thanh, phụ âm, vần) và các luật kết hợp chuẩn cho các cấu phần nội bộ của tiếng.[10] Ví dụ, chỉ có hai thanh sắc và nặng xuất hiện khi các vần tận cùng bằng –p, –t, –c và –ch.

8. Các mẫu cấu tạo trên và các luật kết hợp cho ta khoảng 15.000 tiếng nói được và nhận biết được là tiếng Việt, nhưng chỉ có trên dưới 7.000 tiếng Việt hiện đại dùng đến.

D. KẾT LUẬN

Định nghĩa đơn vị chính tả chính xác hơn định nghĩa ký tự của Unicode (không làm rõ sự tương ứng của chữ biểu ý, gốc ấn, gốc hồi,… và chữ latinh). Nó phát huy lợi thế của Unicode giúp chúng ta làm được tập mã đa ngữ Việt Nam, nằm trong tập mã đa ngữ quốc tế. Kỹ thuật dấu rời (*combining marks*)—những đơn vị chính tả—cho phép chúng ta tiệm cận đặc thù của các thứ chữ viết và tiếng nói trong nước. Nó cho phép chúng ta ghi lại, và nhái lại đúng hoạt động đặc thù của tiếng Việt và các thứ tiếng khác, như nhập dữ liệu (theo phong cách riêng của mỗi thứ chữ viết), sắp thứ tự, tìm kiếm, chuẩn chính tả tự động, chuyển ngữ (ví dụ, Nôm–quốc ngữ và ngược lại), dịch/trữ/tìm/phát sinh âm thanh, sản sinh các cách nói lái, từ láy, vần điệu trong lời

[10] Ngô Thanh Nhàn, *sách đã dẫn*, Chương Ba, *Những nhận xét về mô tả âm vị học tiếng Việt*, trang 59-128.

nói, nhạc, thi ca, v.v. Chúng ta có mục tiêu rộng hơn để làm dễ việc chuyển hoá giữa chữ viết và âm thanh của một thứ tiếng. Trong đó, theo những tri thức/nhận xét về tiếng của tiền nhân, ta gộp các ký tự thành đơn vị lớn hơn, đó là phụ âm, vần và thanh.

Định nghĩa này cho phép hai ngành công nghệ tin học về âm thanh và chữ viết phát triển song song, dành chỗ cho các nhà tin học trẻ tham gia giải quyết quan hệ của chữ viết và tiếng nói, góp phần vào việc tự động thu tin tức đủ loại (vừa tiếng vừa chữ), giúp cho người điếc, người câm, người ngoại quốc du lịch, giảng dạy tiếng Việt tự động, thu thập tri thức (tiềm tàng trong chữ viết và tiếng nói), v.v.