NHÁP

XÂY DỰNG CƠ SỞ TRI THỨC CHỮ NHIỀU BẬC ĐỆ QUY VÀ KHO THÀNH TỐ CƠ BẢN CỦA CHỮ NÔM

Ngô Thanh Giang & Tô Trọng Đức
Ngô Thanh Nhàn & Ngô Trung Việt
Nhóm Nôm Na, Hà Nội

Hội nghị Quốc tế về chữ Nôm
Tháng 6 năm 2006, Huế

Giới thiệu

Chữ Hán-Nôm hiện nay được sử dụng rộng rãi trong vùng Đông Á và trên thế giới nhờ sự phát triển của ngành công nghệ thông tin, mạng Internet và nhất là chuẩn mã chữ quốc tế Unicode và ISO/IEC 10646. Chữ Nôm nhờ đó đã gia nhập cộng đồng mạng thông tin và máy tính.

Tuy nhiên, thông lệ quốc tế về chữ Hán-Nôm còn nhiều điểm cơ bản chưa chính xác về tự đạng. Cụ thể, mỗi chữ Hán-Nôm bị coi là một chữ "cái" (*character*), và từ đó cách phân tích nội tại của chữ Hán-Nôm còn phải dùng phương pháp bộ và số nét theo cách của *Tự điển Khang Hi* năm 1710-1716. Do đó, kho chữ "cái" Hán-Nôm trong bộ chuẩn quốc tế lên đến hơn 50.000.[1] Đó là một điều kỳ lạ. Ai cũng biết mỗi chữ Hán-Nôm ghi một âm tiết, được tạo thành bằng những bộ phận giống nhau về hình dáng. *Tự điển Khang Hi* bắt đầu công tác phân tích và tìm ra 214 bộ (mà phương Tây dịch sai thành *radical*). *Tự điển Khang Hi* có thể coi là một bước cách mạng về mặt phân tích chữ Hán theo các bộ phận tự dạng nội tại của chữ, nó cho phép người ta sắp thứ tự vào một bảng (tự điển) theo một quy trình mà ai cũng truy tìm được. Nhưng việc dùng cách đếm số nét (không phải là bộ phận tự dạng nội tại) làm phức tạp thêm cho việc tìm chữ trong văn bản hay tự điển—không một người thành thạo chữ Hán-Nôm khi nhìn mặt chữ lại nghĩ đến số nét.

Trong bài viết này chúng tôi bàn tới quy trình xây dựng và thống nhất hóa cơ sở tri thức chữ Hán Nôm (sau đây gọi là CSTTC). Gọi là cơ sở tri thức là vì, ngoài việc là kho chữ tập hợp 20.213 chữ Nôm với 37.714 mục từ các nguồn khác nhau, CSTTC Hán Nôm còn lưu giữ các thông tin tự đạng hữu ích cho các thao tác công nghệ thông tin, ngôn ngữ học (từ vựng lịch sử, từ vựng học, ngữ nghĩa học), văn bản học, giải nghĩa Việt-Anh, v.v. Việc thống nhất và hoàn thiện CSTTC được tiến hành trên một quy trình mới: Thành tố với tư cách là các thành phần cấu tạo theo từng bậc cho tự dạng chữ Hán-Nôm.

---

[1] Cho đến nay người ta đã tìm ra khoảng 5.000 chữ Giáp cốt văn nhưng có lẽ còn nhiều chữ chưa tìm ra. Tự điển Đông Hán, *Shuowen jiezi*, do Xu Shen soạn, có 9.353 chữ. *Khang Hy tự điển* soạn trong thời nhà Thanh có 46.964 chữ. *Hán ngữ đại tự điển*, do Nhóm nhà xuất bàn Hubei tỉnh Sichuan năm 1986, có hơn 56.000 chữ.

Quy trình xây dựng và hoàn thiện CSTTC là một quy trình nhỏ của quy trình Nôm na. Nó có quan hệ chặt chẽ với các quy trình con khác trong hệ thống. Quy trình Nôm Na được mô tả như sự tích hợp của các quy trình con sau:

- Tập hợp và xây dựng *cơ sở dữ liệu* thống nhất chữ Hán Nôm;
- Xây dựng cơ sở tri thức mỗi chữ Hán Nôm;
- Xây dựng các công cụ tra cứu – nghiên cứu chữ Hán Nôm;
- Xây dựng bàn phím chữ Hán Nôm;
- Xây dựng kho văn bản chữ Hán Nôm;
- Xây dựng chương trình học tập điện tử cho chữ Hán Nôm (*Nôm E-learning*).

Bài này trình bày ý nghĩa của riêng quy trình xây dựng và hoàn thiện CSTTC mà nhóm Nôm Na đã thực hiện trong thời gian qua, cụ thể là phân tích thành tố[2] theo tự dạng, thiết lập quá trình tạo tự dạng chữ, nhưng không theo lịch sử (dị đại) tạo chữ như các nhà nghiên cứu Hán Nôm hay ngôn ngữ học vẫn dùng. Trên cơ sở đó, đúc kết và khái quát hoá để có thể xây dựng bộ thành tố cơ bản, tiến tới việc xây dựng bàn phím chữ Hán Nôm, và kết quả của quy trình sẽ trở thành một hỗ trợ đắc lực cho việc biên soạn nội dung cho chương trình *Nôm E-learning*.

a. Thành tố là gì?

Thành tố là một bộ phận của chữ Hán-Nôm có nghĩa, là một chữ hay một bộ tạo thành chữ mới. Thành tố có thể được tạo ra bằng các thành tố nhỏ hơn. Thành tố nhỏ nhất không còn phân tích được nữa gọi là thành tố cơ bản. Ở đây chúng tôi chỉ chú ý đến tự đạng của chữ và thành tố. Từ "có nghĩa" gồm có tự đạng xuất hiện trên nhiều chữ khác nhau, và có tên gọi. Tên gọi của thành tố nhiều khi là "âm đọc" của thành tố đó.

Trong bài này, chúng tôi trình bày quy trình Nôm Na, xây dựng chức năng đệ quy vào kho thành tố cơ bản của Nôm na ***dựa trên giả định phân tích nhị phân và cấu tạo nhị phân***. Một ví dụ đơn giản trong truyện dân gian cho thấy cấu tạo nhị phân và đệ quy (nhiều tầng) của chữ:

八 刀 分 米 粉 *bát đao phân mễ phấn*
千 里 重 金 鍾 *thiên lý trọng kim chung*

trong đó quá trình tạo chữ 粉 *phấn* và 鍾 *chung* gồm hai bậc, mỗi bậc có hai chữ nhập thành một:

Bậc 1: 八 + 刀 → 分 và sau đó, bậc 2: 分 + 米 → 粉
Bậc 1: 千 + 里 → 重 và sau đó, bậc 2: 重 + 金 → 鍾

Hai câu đối trên rút ra từ một câu chuyện dân gian duyên dáng và thông minh, tuy cách phân tích quá trình tạo chữ không thật chỉnh. Ví dụ khác rõ hơn, như hai tầng phân tích chữ 唼 *lời*:

Bậc 1: 唼 *lời* → 口 *khẩu* + 歰 *trời*

[2] Chúng tôi dùng chữ "thành tố" gần nghĩa với Lê Văn Quán 1981 nhưng không đi vào lịch sử xuất hiện, cấu tạo hay âm đọc (ngữ âm lịch sử).

Bậc 2: 𡗶 *trời* → 天 *thiên* + 上 *thượng*.

Các ví dụ trên cho chúng ta:

Giả định 1: Thành tố của chữ Hán Nôm là một bộ phận tự dạng có nghĩa của phân tích đệ quy nhị phân của kho chữ.

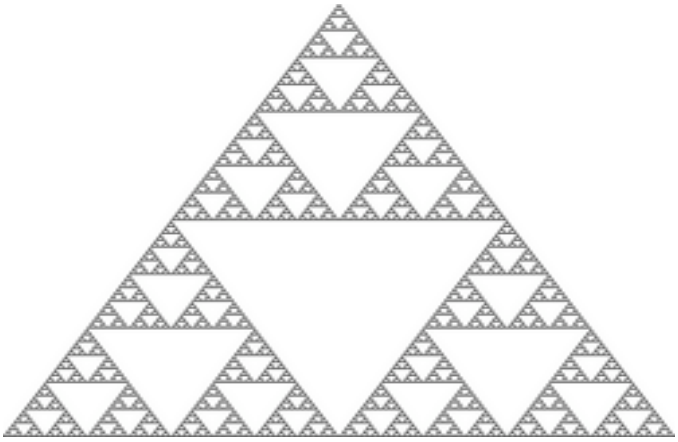
Ta nói, cách đánh vần chữ Hán Nôm của người Việt Nam cho ta hình dung các bộ phận cấu tạo chữ. Chữ do chữ tạo thành, cũng như từ do từ tạo thành.

Giả định 2: Mỗi thành tố là một chữ trong kho, có một mã Unicode duy nhất, có tự dạng và có tên gọi (âm đọc).

Quy trình Nôm Na mất 3 năm tiến hành phân tích nhị phân cho từng chữ trong kho CSTTC, và bài này báo cáo kết quả của quy trình hai giả định trên. Kết quả gồm hai phần: phần theo đúng phân tích nhị phân đệ quy, kèm theo bảng thành tố cơ bản nhất, và phần ngoại lệ.

b. Đệ quy là gì?

Đệ quy (*recursion*) là một thuật ngữ tin học trong lập trình máy tính mô tả các hiện tượng tự nhiên, ngôn ngữ học và toán học. Đây là một lệnh của chương trình làm cho một *modul* (thao tác) hoặc chương trình con tự gọi lại chính mình. Chức năng đệ quy được dùng để bổ sung các sách lược tìm kiếm và thực hiện sắp xếp nội bộ chẳng hạn, trong đó số lượng các lời gọi đệ quy không thể dự đoán được. Cấu trúc của một chữ Nôm gần giống như minh họa trong tam giác Sierpinski. Một chữ Nôm có thể phân tích thành các thành tố bậc 1, bậc 2,… cho tới bậc *n* (bậc tối giản). Bản thân các thành tố ở bất cứ bậc nào đều có thể đã xuất hiện ở đâu đó trong CSTTC.



Tam giác Sierpinski, biểu diễn khái niệm đệ quy

c. Quy trình Nôm Na: xây dựng chức năng đệ quy và kho thành tố cơ bản cho CSTTC

Quy trình Nôm Na là một quy trình đưa các tài liệu Hán-Nôm ròng vào cơ sở tri thức chữ và bộ phông Hán-Nôm nhằm sử dụng đại trà trên mạng internet.

Cơ sở tri thức chữ Hán Nôm chứa thông tin về từng chữ. CSTTC khác với cơ sở dữ liệu ở chỗ nó bao gồm các thông tin liên quan đến công nghệ trao đổi và hiển thị (các loại mã chữ, mã bộ, in ấn, trình bày, sắp thứ tự theo các loại tiêu chí, truy cập,…), thông tin về từ vựng, xuất xứ, thông

tin cấu tạo, thông tin đối chiếu Việt-Anh. Phiên bản sử dụng cho bài viết là 1.07 của CSTTC NomnaTongLight_kB. Chúng tôi chú ý nghiên cứu và tiến hành thao tác trên một số các trường thông tin sau:

1. **ID** (số thứ tự): chỉ báo về trật tự thời gian theo đó các bản ghi được nhập vào. Trường ID là trường khoá để giữ đúng trật tự các bản ghi, để sau khi tiến hành các thao tác xử lý, dựa vào thông tin về ID của chữ, ta có thể tìm lại được trật tự cũ của CSTTC.
2. **Mã Unicode:** ghi lại thông tin về mã Unicode của các chữ Hán-Nôm đã được tổ chức Unicode cấp mã; và các mã thuộc mặt phẳng 6 (60000-6ffff) được cấp cho các chữ Nôm mới trong quy trình Nôm Na, các mã này chưa có trong chuẩn quốc tế Unicode.
3. **Nôm:** chứa hình chữ đại diện của các mã chính thức được thừa nhận, là hình chữ thuộc bộ font Tống thể mảnh NomnaTonglight.ttf.
4. **Quốc ngữ:** chứa thông tin về âm đọc quốc ngữ của hình chữ ở trường **Nôm**.
5. **Mẫu ghép:** gồm các mã ghép từ 2ff0 đến 2ffb, quản lý cách kết hợp của các thành tố, là thể hiện của cách thức cấu tạo chữ.
6. **Hình mẫu ghép:** một trong 12 cách kết hợp thành tố được trực quan hoá.
7. **Thành tố 1:** chứa thông tin về tự dạng của thành tố đầu.
8. **Thành tố 1 – id:** chỉ báo của thành tố - thể hiện chức năng đệ quy của CSTTC
9. **Thành tố 1 – qn:** âm đọc quốc ngữ của thành tố
10. **Mã của thành tố 1:** mã nội bộ quản lý thành tố - trước được dùng căn cứ vào bộ font yếu tố cơ bản của Đỗ Quốc Bảo
11. **Thành tố 2:** chứa thông tin về tự dạng của thành tố thứ hai
12. **Thành tố 2 – id:** chỉ báo của thành tố - thể hiện chức năng đệ quy của CSTTC
13. **Thành tố 2 – qn:** âm đọc quốc ngữ của thành tố
14. **Mã của thành tố 2:** mã nội bộ quản lý thành tố - trước được dung căn cứ vào bộ font yếu tố cơ bản của Đỗ Quốc Bảo
15. **Thành tố 3:** chứa thông tin về tự dạng của thành tố thứ ba
16. **Thành tố 3 – id:** chỉ báo của thành tố - thể hiện chức năng đệ quy của CSTTC
17. **Thành tố 3 – qn:** âm đọc quốc ngữ của thành tố
18. **Mã của thành tố 3:** mã nội bộ quản lý thành tố - trước được dung căn cứ vào bộ font yếu tố cơ bản của Đỗ Quốc Bảo
19. **Bộ thủ** (*Radical*): chứa thông tin về tự dạng của bộ thủ
20. **Bộ thủ – qn**: âm đọc quốc ngữ của bộ thủ
21. **Mã bộ URN** (Unicode Radical Number): Mã bộ Unicode của bộ thủ[3]
22. **Sunicode:** Số nét còn lại của chữ theo Unicode.
23. **KTotalStrokes:** Tổng số nét của chữ, kể cả số nét của bộ thủ.

NomnaTongLight_kB phiên bản 1.07 gồm có 37.714 mục (*record*), mỗi mục là một tập hợp con các tri thức về một tự dạng và một âm đọc quốc ngữ. Thông thường, thông tin về điểm mã (*codepoint*) là chỉ báo quan trọng nhất để phân biệt các chữ: mỗi tự dạng có một điểm mã quốc tế duy nhất. Do đó, khi làm việc trên CSTTC, chúng tôi chủ yếu căn cứ vào trường ID, và điểm mã của chữ.

[3] Xem danh sách bộ Unicode của Nôm Na tại http://nomfoundation.org/radicals.html.

Tuy bài này chỉ chú ý đến phân tích nhị phân, nghĩa là mỗi chữ chỉ chứa nhiều nhất là hai thành tố, chúng tôi vẫn dành chỗ cho khu vực Thành tố 3 trong CSTTC, cho khả năng phân tích tam phân, tuy danh sách này nhỏ. Xem danh sách 3 thành phần giống nhau kèm theo.

Vì quy trình chủ vào việc phân tích thành tố theo tự dạng chữ, nghĩa là phân tích kho chữ, nên việc đầu tiên là lọc bớt các trường hợp trùng điểm mã là thao tác cần thiết trên CSTTC—theo **Giả định 2** ở trên. Sau khi lọc bớt các mục từ trùng điểm mã, CSTTC còn lại 20,213 mục từ. Nói cách khác, kho Nôm Na hiện có 20.213 chữ, hay 20.213 điểm mã, duy nhất.

Các công việc cần thực hiện trên CSTTC bao gồm:

- Thống nhất CSTTC: thống nhất tên bộ, số URN; thống nhất tự dạng của các thành tố và kiểm tra chính tả cho thành tố và tên bộ.
- Xây dựng chức năng đệ quy cho CSTTC trên cơ sở một kho chữ đã thống nhất về tên gọi (âm đọc quốc ngữ) và chuẩn chính tả.
- Xây dựng tập hợp thành tố cơ bản dựa trên CSTTC đệ quy.

I. Thống nhất CSTTC

1. Thống nhất trường thông tin về bộ và mã bộ (URN)

Mã bộ *(Unicode Radical Number* hay URN) là số thứ tự của bộ thủ theo trật tự của Unicode (hay Khang Hi mở rộng). Mỗi một mã được gán cho bộ thủ theo trật tự số nét của bộ thủ, tương ứng với thứ tự bộ thủ trong *Tự điển Khang Hi*. Như vậy, giữa bộ thủ (*radical*) và mã bộ URN có sự tương ứng.

Trên CSTTC, tham chiếu với Bảng bộ thủ [*Unicode Radical List*] ta có thể tìm ra những bản ghi cùng mã nhưng có trường Radical và URN không trùng khớp. Từ đó sửa lại thông tin về trường Bộ thủ (Radical) và Mã bộ URN cho đồng nhất. Đây là công tác liên tục, bán tự động, nhằm tìm ra lỗi và không nhất quán trong một kho chữ ngày càng lớn có nhiều chữ Hán Nôm có tự dạng giống nhau nhưng tránh trường hợp có mã khác nhau.

> **Giả định 3:** Hai chữ Hán Nôm giống nhau phải có cùng bộ và số nét.

Trên đây có thể gọi là giả định đương nhiên [*default*]. Vì chúng tôi chỉ thao tác trên tự dạng, nên kết quả có thể khác với lịch sử tạo chữ.

Có khi có chữ Nôm có lịch sử cấu tạo khác với chữ Hán cùng tự dạng, có phân tích bộ hay thành tố khác nhau. Khi phân tích thành tố, thành tố có cùng tự dạng, vì mỗi thành tố là một chữ, nên có thể có nhiều hơn một "tên gọi" (cách đọc). Ví dụ:

a. 垮 *khoai* và *khoa*, theo Vũ Văn Kính 1971 có quá trình tạo chữ khác nhau:

   垮 *khoa* (HV) → 土 *thổ* + 夸 *khoa* ("sụp đổ, phá đổ")

   垮 *khoai* (Nôm) → 土 *thổ* + ½ chữ 侉 *khoa*.

   Vậy, 夸 *khoa* hay ½ chữ 侉 *khoa* (hay theo phân tích của Lê Văn Quán là bỏ bớt bộ thủ) về tự dạng chỉ là một thành tố.

b. 猥 *voi*, *vỏi* (Nôm) và wei4 "a kind of beast, a legendary monster"

猥 *vệ* (HV) → 犭 *khuyển* + 爲 *vi*.

猥 *voi*, *vỏi* (Nôm) → 犭 *khuyển* + 爲 *vay, vây, ve, veo, vi, vị, vì, vơ, vờ, vời, với.*

c. 魦 sa (có khi viết 鯊 sa) theo Lê Văn Quán (tr. 83)

魦 *sa* (HV) → 魚 *ngư* + 少 *thiếu*

魦 *sa* (HV) → 魚 *ngư* + 少 *sa* (½ chữ 沙 *sa*).

Vậy, 少 *thiếu* hay ½ chữ 沙 *sa*, về tự dạng chỉ là một thành tố, có hai âm đọc, *thiếu* và *sa*. Tương tự, danh sách các chữ có một thành tố bị "bỏ bớt bộ thủ" của Lê Văn Quán (trang 91) gồm:

Chữ Nôm	**Ghi ý**	**Ghi âm**
luộc 燘	火 *hoả*	录 (綠) *lục*
lóc 鯥	魚 *ngư*	录 (祿) *lộc*
khê 糓	米 *mễ*	奚 (溪) *khê*
chửa 媎	女 *nữ*	者 (渚) *chử*
dặm 黕	里 *lý*	炎 (淡) *đạm*
chưa 褚	未 *vị*	者 (渚) *chư*
uống 吂	口 *khẩu*	王 (汪) *uông*
húi 剞	刂 *đao*	每 (悔) *hối*
hỏi 嗨	口 *khẩu*	每 (悔) *hối*
đất 坦	土 *thổ*	旦 (怛) *đát*

Các thành tố 录 *lục/lộc,* 奚 *khê/hề,* 者 *giả/chử (dã, trả),* 炎 *viêm/đạm,* 王 *ngọc/uông (vương, vướng),* 每 *mỗi/hồi (mỏi, mọi, môi, mỗi, mối, mũi, múi, muối),* 旦 *đán/đát (chán, dán, đáng, đắn, đến, trán),…*

d. 能 *năng* thuộc bộ 月 *nhục*, trong khi chữ tắt của nó là 𠂤 *năng* thuộc bộ 匕 *chuỷ*? Quá trình viết tắt sản sinh ra những chữ mới có thể biến thành các bộ khác nhau, hay các thành tố khác nhau. Ví dụ: 飞 viết tắt của 飛, 疋 *sơ* viết tắt của 疋 (hay theo Lê Văn Quán là chữ 歷 *lịch*),… nay đã thành bộ mới trong UniHan (gọi là bộ phụ gia).

2. Thống nhất tên gọi thành tố

Chữ Nôm được cấu tạo từ những thành phần sau:

1. Thành phần tham gia cấu tạo có nguồn gốc từ bộ phận chữ Hán, thành phần này thường là thành phần có nghĩa, có thể đứng độc lập. Ta gọi đây là một thành tố. Các thành phần tham gia cấu tạo chữ Nôm với tư cách là các bộ, các chữ Hán, hoặc các chữ Nôm vốn là các thành phần có nghĩa và có tên gọi. Tên thành tố chữ Hán được kiểm tra và đối chiếu với các nguồn:

— Các từ/tự điển của Vũ Văn Kính, Nguyễn Kim Thản, Hồ Lê, Trần Văn Kiệm, Trương Đình Tín, v.v.
— Unicode UniHan Database: http://www.unicode.org/charts/unihan.html

Âm Hán Việt được sử dụng làm tên gọi của thành tố chữ Hán, căn cứ trên Bính Âm của tiếng Hán, nếu từ điển Hán Việt của Việt Nam chưa có.

2. Thành phần tham gia cấu tạo với tư cách là bộ phận thuần Nôm:
   – Chữ viết tắt theo chữ Hán Nôm có sẵn
   – Chữ Nôm tham gia như một thành phần cấu tạo một chữ Nôm khác.

   Đối với thành phần tham gia cấu tạo là các chữ Nôm, âm quốc ngữ (được hiểu như cách đọc nôm na) được dùng làm tên thành tố.

3. Thành phần tham gia cấu tạo chưa có tên:
   Thành phần cấu tạo biểu âm theo loạt vẫn được coi là thành tố. Đối với các thành phần cấu tạo biểu âm theo loạt, tên thành tố được gán như tên của một chữ có mật độ xuất hiện lớn trong các văn bản.

— Các chữ có thành tố 𠕻 U+2057B trong các chữ sau đây mà Lm Trần Văn Kiệm cho là viết tắt của 堯 *nghiêu*.

xuống	*mồng*	*bay*	*quan, quán*
[illegible]	[illegible]	[illegible] [illegible] [illegible] [illegible]	冠

Chữ 𠕻 U+2057B trong kho UniHan không có cách đọc (chỉ có trong *Hán Ngữ đại tự điển*), nhưng có thể có cấu tạo 冖 *mịch* và 元 *nguyên*. Chúng tôi chọn dùng cách đọc, *nghiêu*.

— Các chữ có thành tố 咅 pou4 U+5485 ‘to spit out’ trong các chữ:

bội, bòi	*bộ*	*bồ*	*bội, bụi*	*bồi*	*bội, vùi*
培	部	菩	倍	錇	焙

Những âm *bội, bòi, bộ, bồ, bồi, vùi*, đều có phụ âm đầu là môi hữu thanh /b/ và /v/, có âm cuối là bán nguyên âm /i/ hay mở, có nguyên âm trung tâm sau tròn môi /ô/ hay /u/, có thanh thấp (*low register* tương ứng với hữu thanh) huyền hay nặng. Do đó ta có thể tái lập *bồi*. Cụ Vũ Văn Kính đưa ý kiến đọc là “*nửa chữ bội*”, “*nửa chữ bồi*” (*Học chữ Nôm*, trang 46).

— Các chữ có thành tố 畐 fu2 U+7550 “to fill; fold a cloth” có nghĩa và có âm đọc. Lm Trần Văn Kiệm và Vũ Văn Kính cho các chữ Nôm 愊 煏 堛 踾 là “nửa chữ *bức*”. Do

chữ 𤾓 *bức* thuần Nôm, ta có thể cho hai cách đọc, *phúc* theo chữ Hán hay *bức* theo chữ Nôm.

phúc	*bậc*	*bặc – bặt*	*bức – bực*	*bậc*
幅 愊 煸 福 輻	堛	匐	幅 湢 蝠 偪 逼 踾	簏

— Các chữ có thành tố 冓 *cấu* U+5193.

cấu	*cáu, gẫu, quạu*	*giảng, nhãng*	*quảu*	*cấu*	*cấu*	*gấu*	*bấu*	*cấu*	*cấu*	*cấu*
傋	嗙	講	篝	媾	購	褠	搆	構	覯	遘

– Thành phần cấu tạo vô nghĩa, vô thanh vẫn được coi là thành tố.
Đối với các thành phần cấu tạo vô nghĩa và vô thanh, tên thành tố được đánh dấu bằng kí hiệu "n/a" (có nghĩa là tạm thời chưa có tên). Việc đặt tên cho các thành tố này sẽ tuân theo quy tắc của ngữ âm học lịch sử, phương pháp như trên.

II. Xây dựng CSTTC đệ quy trên cơ sở tên các thành tố đã được thống nhất

Đặc tính của CSTTC thích hợp cho mô tả đệ quy khi được xây dựng theo mô hình Backus Naur Form như mô tả các trường ở trên. Trong mô tả của từng mục tự gồm có:

1. Mục có hai thành tố là *nhánh* trong quy trình đệ quy;
2. Mục hai thành tố đều trống là thành tố cơ bản, là *lá* trong tiến trình đệ quy.
3. Mục có thành tố ghi "n/a" là nhánh chưa biết cách xử lý (phân tích).

Từ đó, chúng ta có thể:

Rút ra quá trình cấu tạo của một chữ (có người gọi là *tự nguyên*) bằng cách dùng quy trình đệ quy theo nhánh đi sâu trước (*depth-first*), từ trái sang phải (*left-to-right*) cho đến khi chạm hết *lá*.

Unicode	**Nôm**	**QN**	**Mẫu**	**tt1**	**tt1_qn**	**tt2**	**tt2_qn**	**Bộ**	**Bộ_qn**	**URN**	**Nét**
20CD2	哱	lời	⿰	口	khẩu	孛	lời	口	khẩu	0030	7
53E3	口	khẩu						口	khẩu	0030	0
215F6	𡈼	giời	⿱	天	thiên	上	thượng	大	đại	0037	4
215F6	𡈼	lời	⿱	天	thiên	上	thượng	大	đại	0037	4
215F6	𡈼	trời	⿱	天	thiên	上	thượng	大	đại	0037	4
5929	天	thiên	⿱	一	nhất	大	đại	大	đại	0037	1
5929	天	thiêng	⿱	一	nhất	大	đại	大	đại	0037	1

5929	天	thiên	⿱	一	nhất	大	đại	大	đại	0037	1
4E0A	上	thượng	⿱	卜	bốc	一	nhất	一	nhất	0001	2
5927	大	đại						大	đại	0037	0
5927	大	dảy						大	đại	0037	0
5927	大	dãy						大	đại	0037	0
5927	大	đai						大	đại	0037	0
4E00	一	nhất						一	nhất	0001	0
4E00	一	nhắt						一	nhất	0001	0
4E00	一	nhứt						一	nhất	0001	0
2E8A	卜	bốc						卜	bốc	0025	0
5171	共	cọng						八	bát	0012	4
5171	共	cộng						八	bát	0012	4
5171	共	cùng						八	bát	0012	4
5171	共	cũng						八	bát	0012	4
5171	共	cụng						八	bát	0012	4
5171	共	gọng						八	bát	0012	4
5171	共	cộng						八	bát	0012	4
20017	丼	khệnh						一	nhất	0001	4
20016	丱	khạng						一	nhất	0001	4
…	…	…									

Theo bảng trên:

- Mỗi hàng đều có cột Unicode, Nôm và quốc ngữ, trong đó cột Nôm và cột điểm mã Unicode là tên của hàng.
- Hàng của chữ là bộ thủ có: ô Nôm = ô Bộ, ô QN = ô Bộ_qn, ô Nét = 0;
- Hàng thành tố cơ bản có các ô Mẫu, tt1, tt1_qn, tt2 và tt2_qn trống.

Tiếp tục quy trình cho tới khi không mọi thành tố đều được quy về tối giản.

Để tìm quá trình cấu tạo một chữ, ta chỉ cần:

1. Tìm tất cả các hàng có tự dạng chữ muốn tìm trong cột **Nôm** của CSTTC. Ví dụ, trong CSTTC trên, 哽 có 1 hàng, 大 có 4 hàng, 天 có 3 hàng, 共 có 7 hàng, v.v.
   Chọn 1 hàng,
   a. Nếu các ô **Mẫu**, **tt1**, **tt1_qn**, **tt2** và **tt2_qn** trống, đây là một thành tố cơ bản;
      i. nếu các ô **Nôm** = ô **Bộ**, ô **QN** = ô **Bộ_qn**, ô **Nét** = 0, đây là thành tố cơ bản và là một bộ thủ Unicode;
      ii. hết (nhánh đang tìm).
   b. Ngược lại, nếu các ô **Mẫu**, **tt1**, **tt1_qn**, **tt2** và **tt2_qn** không trống, làm 2 động tác:
      i. Tìm thành tố **tt1** theo 1)
      ii. Tìm thành tố **tt2** theo 1)

2. hết (nhánh đang tìm, CSTTC không có chữ muốn tìm).

Quy trình đệ quy xảy ra liên tiếp tại vị trí 1.b., theo nhánh **tt1** trước (trái trước, và sâu trước). Khi hoàn toàn xong **tt1**, mới sang **tt2**. Thiết lập chức năng đệ quy cho CSTTC là việc cần thiết cho quy trình tổ chức một bàn phím, truy cập chữ Hán Nôm trực tiếp hơn, không thông qua chữ quốc ngữ hay bộ và số nét.

III. Xây dựng bảng thành tố cơ bản của NomnaTongLight.ttf

Áp dụng nguyên lý tạo chữ bằng cách lắp ghép các thành tố cơ bản theo các biến thể vị trí, bảng thành tố cơ bản của NomnaTongLight.ttf được xây dựng trước tiên là trên kết quả của việc xây dựng chức năng đệ quy cho CSTTC. Thành tố cơ bản của NomnaTongLight.ttf là tập hợp các thành tố tối giản và các biến thể vị trí của chúng.

Bảng thành tố cơ bản Nôm Na chứa 349 thành tố cơ bản (xem bảng đính kèm) gồm có bộ và thành tố không phân tích nhị phân được nữa. Chúng tạo ra 19.811 chữ (số chữ trong kho NomNaTong Light ttf), và 406 hàng ngoại lệ.

Giả định 4: Một chữ không tìm ra cách phân tích nhị phân có thể tạm coi là một thành tố cơ bản, chờ giải pháp phụ gia.

Một ví dụ trong các trường hợp ngoại lệ, có các chữ gồm 3 thành tố giống nhau xếp theo hình tam giác. Có chữ giải quyết được bằng phương pháp đệ quy nhị phân cấu tạo trên-dưới ⿱ nhờ chữ đã có gồm hai thành tố giống nhau, như sau:

人 *nhân* + 从 *tòng, tùng, tụng* → 众 *chúng*	女 *nữ* + 奻 → 姦 *gian*
子 *tử* + 孖 → 孨	手 *thủ* + 拜 → 掱 *vát (thủ)*
日 *nhật* + 昍 → 晶 *tinh*	木 *mộc* + 林 *lâm* → 森 *sâm*
水 *thuỷ* + 沝 → 淼 *diễu, miểu*	火 *hoả* + 炏 → 焱 *yên, diễm*
牛 *ngưu* + 牪 → 犇 *bôn*	田 *điền* + 畕 → 畾 *lei4, điệp*
石 *thạch* + 砳 → 磊 *lỗi, luộc*	耳 *nhĩ* + 聑 → 聶 *nhiếp*
虫 *trùng* + **蚰** *côn* → 蟲 *trùng*	言 *ngôn* + 誩 → 譶
貝 *bối* + 賏 → 贔 *bí*	金 *kim* + 鍂 → 鑫 *hâm*
隹 *chuy* + 雔 → 雥	馬 *mã* + 駜 → 驫
龍 *long* + 龖 → 龘	v.v.

Các chữ gồm 3 thành tố giống nhau nhưng không thể phân thành mẫu ghép chuẩn như 品 *phẩm,* 垚, 尛, 惢 *nhị,* 歮, 毳 *thuế, thuý*, 皛, 𥆞, 羴 *thiên*, 舙, 芔, 轟 *oanh,* 飍, 馫, 鱻 *tiên*, 矗 *súc*, 燊, [illegible], [illegible], v.v. tuy không nhiều, nhưng là một hệ ngoại lệ có hệ thống, cho quy trình đệ quy nhị phân.

KẾT LUẬN

Quy trình Nôm na: chữ Nôm trên mạng là một trong những dự án được Hội Bảo tồn Di sản chữ Nôm tiến hành tổ chức xây dựng, nhằm mang lại diện mạo mới, cách nhìn mới về chữ Nôm Việt Nam. Cách phân tích đệ quy nhị phân chữ Hán-Nôm trong kho CSTTC Nôm Na theo phương pháp truyền thống *trên trước dưới sau, ngoài trước trong sau, trái trước phải sau*, cho chúng ta một kết quả ban đầu khích lệ với chỉ có 349 thành tố cơ bản, giải thích 98% kho chữ Hán-Nôm. Nó cho phép chúng ta hình dung một quy trỉnh xây dựng bàn phím trực tiếp, thay vì thông qua chữ quốc ngữ như hiện nay.

Tại Việt Nam nhóm Nôm Na đã tiến hành xây dựng quy trình làm phông chữ Nôm. Bước đầu nhóm đã thực hiện thành công việc tạo phông chữ Hán−Nôm với kho chữ gồm 20.213 chữ Hán Nôm. Những bước tiếp theo dự định sẽ chế tạo các phông cho chữ Hán−Nôm theo các thể loại khác nhau, có nguồn gốc xuất xứ từ những văn bản Nôm tiêu biểu cổ nhất cho đến những văn bản mới nhất, đáp ứng được những nhu cầu làm công tác chế bản cũng như in ấn các văn bản Nôm khác nhau, nhằm khôi phục lại nguyên bản những tác phẩm Nôm bằng công nghệ thông tin hiện đại.

Với việc xây dựng một CSTTC thống nhất, và chức năng đệ quy thể hiện trong CSTTC đó, Nôm Na hy vọng đóng góp những tiện ích mới giúp giản lược đi những công việc bằng tay và phục vụ thiết thực cho công cuộc nghiên cứu cũng như bảo tồn chữ Nôm.

Tham khảo

Các tập mã chữ Nôm do Việt Nam cung cấp cho nhóm ISO/IEC 10646 JTC1/IRG từ năm 1994 đến nay: NPCT 2.1, TCVN 5712: 1993, TCVN 5773: 1993, TCVN 6056: 1995, VHN1: 1998, VHN2: 1998. Đề nghị CJK Extension C1 của Việt Nam.

ISO/IEC 10646 JTC1/IRG từ năm 1994 đến nay, Unihan 3.1 Radical-Stroke Index.

Khang Hi Tự điển. Trung Quốc Cổ điển Tinh phẩm ảnh ấn tập thành: (Thanh) Trương Ngọc thư đẳng biên soạn. Thượng Hải Văn nghệ xuất bản xã. 2000.

Linh mục Trần Văn Kiệm. *Giúp đọc Nôm và Hán Việt*. Nhà xuất bản Đà Nẵng và Hội Bảo tồn Di sản chữ Nôm, 2004.

Vũ Văn Kính. *Học chữ Nôm*, Nhà xuất bản Đồng Nai, 1995.

Ngô Thanh Nhàn, Ngô Trung Việt và Nhóm Nôm Na. *Quy trình Nôm Na*, trình bày *tại Hội thảo Hè* 2002, Đại học Maine.

Viện Ngôn ngữ học. *Bảng tra chữ Nôm*. Nhà xuất bản Khoa học Xã hội, Hà Nội 1976.

Lê Văn Quán. *Nghiên cứu về chữ Nôm*. Nhà xuất bản Khoa học Xã hội. Hà Nội. 1981.

Nguyên Kim Thản, chủ biên. 2000. *Tự điển Hán Việt hiện đại*. Nxb Thế giới, Hà Nội.

Trương Đình Tín. *Bảng Phiên âm Nôm Việt*. Nhà xuất bản Thuận Hóa, 2003.

Nguyễn Quang Xỷ & Vũ Văn Kính. *Tự Điển Chữ Nôm*. Trung tâm Học liệu, Sàigòn 1971.